Exploring Mechanisms of Size Control and Genomic Duplication in *Saccharomyces cerevisiae*

A Computational Systems Biology Study

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr.rer.nat.
im Fach Biophysik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
M.Sc. Thomas Wolfgang Spiesser

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Dr. Andreas Herrmann

Gutachter:
1. Prof. Dr. Dr. h.c. Edda Klipp
2. PD Dr. Martin Falcke
3. Prof. Dr. Andreas Möglich

eingereicht am: 02.08.2011
Tag der mündlichen Prüfung: 19.12.2011

Abstract

One of the most fundamental processes in biology is reproduction, i.e. transmitting genomic information to future generations. To achieve this, single cellular organisms grow, proliferate and divide. The necessary prerequisite for this is acquiring sufficient cellular resources to double size and all cellular components, herein, most importantly the DNA. Defects in either sufficient gain in size or chromosomal doubling can be severe for the organism and has been related to complex diseases in humans, such as cancer. Therefore, the cell has developed sophisticated regulatory mechanisms to control the orderly fashion of growth and duplication.

We have developed mathematical formulations (models) to study systemic properties on different levels of two main cell cycle events, namely size control and DNA replication in the premier eukaryotic model organism *Saccharomyces cerevisiae*. Computer modeling is one part of an interdisciplinary field in science, called systems biology, that combines theoretical and experimental research to provide an integrative view on complex biological systems. Herein, different levels of abstraction e.g. single cell in contrast to population behavior, can open new and different perspectives on a problem which can help understanding the complex nature of dynamic systems.

Along these lines, we have created several models of varying granularity to study cell size homeostasis and genomic duplication. Thus, we provide a single cell model which is based on ordinary differential equations and a stochastic component to explore size control. We deduced population behavior from the single cell model through multi-cell simulations using an environment that we especially developed for this purpose. Also, to study genomic duplication, we implemented an algorithm that simulates the DNA replication process. We used this algorithmic model to test the impact of different replication origin activation patterns. Additionally, we assessed elongation dynamics with a fine-grained stochastic model for the replication machinery motion along the DNA template strand. We complemented our analysis of DNA replication by studying the functional association of genes and replication origins using hypergeometric gene ontology association tests.

Our systems-level analysis reveals novel insights into the coordination of growth and division, namely that (i) size regulation is an intrinsic property of yeast cell populations and that neither signaling nor a size sensing mechanism is required for it, (ii) that DNA replication is robust against perturbations, especially in small chromosomes with high origin density, (iii) that there are distinct locations in the genome where the elongation process is strongly biased and (iv) that catabolic genes are over-represented near early origins and anabolic genes near late origins. Moreover, we provide testable model predictions to guide future experiments and outline follow-up studies for further theoretical analysis to increase systemic understanding of size control and genomic duplication.

The work I present here, explores mechanisms of size control and DNA replication in *Saccharomyces cerevisiae* using an integrative approach to contribute to explaining experimentally observed and not completely understood features of both systems.

Keywords: systems biology, budding yeast, size control, DNA replication, multiscale simulations, ODE model, stochastic model, gene ontology

Zusammenfassung

Ein der Biologie zugrunde liegender Prozess ist die Fortpflanzung, d.h. Weitergabe genetischen Materials an Nachkommen. Einzeller wachsen dazu heran und teilen sich. Grundlage hierfür sind ausreichend Nahrung und Ressourcen, um die eigene Masse und alle Zellbestandteile, insbesondere die DNS, zu verdoppeln. Fehler bei der Wachstumsregulation oder der DNS-Verdopplung können schwerwiegende Folgen haben und stehen beim Menschen im Zusammenhang z.B. mit Krebs. Deshalb haben Zellen Instanzen entwickelt, die den Ablauf von Wachstum und Teilung kontrollieren.

In dieser Arbeit werden mathematische Modelle für die Mechanismen zur Wachstumsregulierung und DNS-Verdopplung in der Bäckerhefe, *Saccharomyces cerevisiae*, vorgestellt. Modellierung ist Teil des interdisziplinären Forschungsfelds Systembiologie, welches theoretische und experimentelle Arbeit kombiniert, um integrative Sichtweisen auf komplexe biologische Systeme zu entwickeln. Hierbei können verschiedene Ebenen der Abstraktion, z.B. das Verhalten einer Zelle im Gegensatz zur Zellkultur, beitragen, neue Betrachtungsweisen zu erschließen und sich damit dem Verstehen komplexer, dynamischer Systeme anzunähern. Wir haben mehrere Modelle für unterschiedliche Ebenen von Wachstum und Teilung entwickelt, u.a. ein Modell für einzelne Zellen, welches auf Differenzialgleichungen basiert. Wir leiten das Wachstumsverhalten von Zellkulturen von diesem Modell ab, indem wir eine Vielzahl von Zellen gleichzeitig simulieren. Dies geschieht mittels einer, von uns speziell zu diesem Zweck entwickelten Software. Außerdem haben wir einen Algorithmus entwickelt, welcher die Möglichkeit bietet, die Verdopplung der DNS zu simulieren. Dieser wurde genutzt, um Auswirkungen verschiedener Aktivierungsmuster auf die Replikation zu testen. Zusätzlich wurde die Verlängerung entstehender DNS Stränge, Elongation, mit einem detaillierten, stochastischen Modell untersucht. Wir haben unsere Ergebnisse zur DNS-Verdopplung mit einer abschließenden Untersuchung ergänzt, die funktionelle Beziehungen von Genen aufzeigt, welche sich in unmittelbarer Nähe zu den Aktivierungsstellen der Verdopplung befinden.

Folgende Einsichten in die komplexe Koordination von Wachstum und Teilung wurden durch den systemorientierten Ansatz gewonnen: (i) Wachstumskontrolle ist eine inhärente Eigenschaft von Hefezellpopulationen, welche weder Signale noch Messmechanismen benötigt, (ii) die Verdopplung des Genoms ist robust gegenüber Störungen, insbesondere in kleinen Chromosomen mit hoher Dichte an Aktivierungsstellen, (iii) Elongation ist über weite Strecken uniform, weicht aber an genau definierten Stellen signifikant ab und (iv) Gene, die für katabole Prozesse kodieren, häufen sich nahe der frühen Aktivierungsstellen und Gene von anabolen Prozessen nahe der späten. Die Modelle sagen das Verhalten beider biologischer Systeme voraus, was unter anderem dazu dient, gezielt Experimente vorzuschlagen, die die Vorhersagen entsprechend überprüfen. Auch werden weiterführende, theoretische Ansätze diskutiert, die das Systemverständnis von Wachstum und Teilung vertiefen könnten.

Die vorliegende Arbeit dient in erster Linie der Erkundung von zellulären Mechanismen zur Wachstumskontrolle und DNS-Verdopplung in *Saccharomyces cerevisiae*, wobei ein integrativer Ansatz dazu beitragen soll, experimentell beobachtete, jedoch bisher nicht vollständig verstandene Eigenschaften beider Systeme zu erklären.

Schlagwörter: Systembiologie, Bäckerhefe, Größenkontrolle, DNS-Verdopplung, Multi-Skalen-Simulation, Gewöhnliche Differenzialgleichung, Stochastisches Modell, Gen-Ontologie

Contents

Abbreviations

Abbreviation	Meaning or Context
AIC	Akaike Information Criterion
A	area
ARS	autonomously replicating sequence
ACS	ARS consensus sequence
bp	base pairs
Cdk1	cyclin dependent kinase Cdc28
CDR	Clb5-dependent-region
non-CDR	non-Clb5-dependent-region
DNA	deoxyribonucleic acid
ECDF	empirical cumulative distribution function
E. coli	*Escherichia coli*
fL	femtolitre
G_1	first gap phase of the cell cycle
G_2	second gap phase of the cell cycle
GO	gene ontology
HH	heavy:heavy
HL	heavy:light
kb	kilo bases
L-BFGS-B	limited-memory Broyden-Fletcher-Goldfarb-Shanno method for bound-constrained optimization
M phase	Mitosis
(m)RNA	(messenger) ribonucleic acid
MSE	multiscale simulation environment
ODE	ordinary differential equation
ORC	origin recognition complex
PDE	partial differential equation
RSS	sum of squared residuals
R^2	coefficient of determination
S phase	the synthesis phase of the cell cycle
S. cerevisiae	*Saccharomyces cerevisiae*
SGD	*Saccharomyces* Genome Database
V	volume

1 Introduction

1.1 Outline

1.1.1 Objective

The most fundamental process in the biology of every living organism is reproduction, i.e. producing healthy descendants. Although the reproduction process differs between species, some basic traits are common to all life forms. These are being born, growing and giving birth in some form of this sense. On the single cell level, this is usually realized in the cell division cycle. The division cycle coordinates all processes required for duplication (Mitchison, 1971). For the unicellular eukaryote budding yeast, *Saccharomyces cerevisiae*, this represents the time from the birth of a cell to the time it splits into two, thereby giving birth to another cell. The cell cycle is characterized by an well-ordered sequence of basic cellular events which divide it into four phases (sketched in Figure 1.1). The first Gap (G_1) phase is the cell cycle stage that is mainly devoted to cell growth and mating. In this phase, the cell must increase adequately in size and metabolic capacity. Furthermore, it has to gather sufficient cellular resources to make a fully functional, reasonably sized, well-equipped cell. G_1 is followed by the Synthesis (S) phase in which, among other things, all genetic information in the form of the deoxyribonucleic acid (DNA) is replicated in order to provide two unique copies that can later on be distributed between mother and daughter cell. The transition from G_1 to S phase is marked by the appearance of a bud, the nascent daughter cell. After completing S phase, cells enter the second Gap (G_2) phase, which is also devoted to growth and to the preparation for cell division. In the last cell cycle phase, Mitosis or M phase, the chromosomes are separated and distributed between mother and daughter cell. When they finally split, both enter a new cell division cycle (Alberts et al., 2007).

The cell division cycle is primarily driven by the sequential accumulation and destruction of cyclins, which act as activators and targeting subunits for the constitutively present cyclin dependent kinase Cdc28 (Cdk1)(Morgan, 1995; Pines, 1995). The active kinase complexes are universal cell cycle regulators conserved from yeast to mammals (Lee and Nurse, 1987). Furthermore, the cell cycle is regulated *via* checkpoint mechanisms. Checkpoints are surveillance systems ensuring that crucial cellular events are completed before the cell enters the next division cycle stage (Hartwell and Weinert, 1989). In this manner, controls are set in place to guarantee that (1) cells only commit to division, if environmental conditions are favorable enough, if cells have attained a critical size and if they have gathered sufficient resources and (2) DNA replication has successfully been completed before chromosomal segregation and cellular division (Alberts et al., 2007). In multicellular organisms deregulation of the cell cycle and its

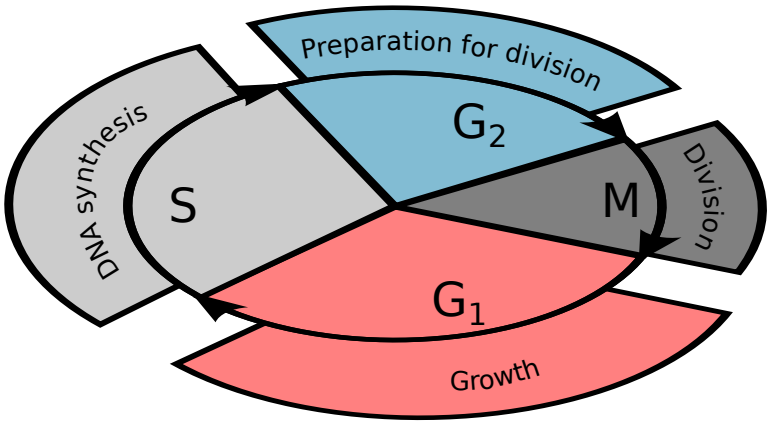


Figure 1.1: **Scheme of the cell division cycle.** The division cycle is divided into four phases, according to relevant cellular events: growth in G_1, DNA synthesis in S, growth and preparation for division in G_2 and division in M.

controls is implicated in the formation of numerous hereditary diseases as well as cancer (Hanahan and Weinberg, 2000). The failure of the checkpoints can lead to (1) abnormal growth and proliferation as a result of unresponsiveness to internal and external growth stimuli, which can be fatal (Hanahan and Weinberg, 2011) and (2) genomic instability, which is an important factor in the formation of cancer (Nurse, 2000). Since the mechanisms of size regulation in G_1 and of DNA replication during S phase are such important aspects of the growth and division cycle, we explore them in greater detail. To this end, we have developed detailed mathematical formulations of both processes for the model organism *S. cerevisiae*. I present those in this thesis.

1.1.2 Scope of the Thesis

The understanding of complex biological processes, which require the interaction of a large number of components in order to function, has strongly been improved by the construction of mathematical models. These models are able to capture the underlying regulatory wirings and predict the dynamics of the process under a variety of conditions (Chen et al., 2000).

In the past and in the present, the cell cycle has been a popular target for mathematical modeling. Detailed deterministic models were used to study robustness and dynamics of the regulatory circuitry of the cell cycle (Chen et al., 2004; Li et al., 2004). Stochastic versions of a toy model, that is based on an earlier model by Tyson and Novak (2001), were used to study the effect of noise on size and cycle time distributions in yeast (Sabouri-Ghomi et al., 2008; Kar et al., 2009; Barik et al., 2010). Furthermore, there are models that focus on specific aspects of the cell cycle, e.g. the G_1/S transition network with respect to cell size at S phase initiation (Barberis et al., 2007). However, most

of these models directly define a critical size, a division ratio, or both, making them unsuitable to study the mechanisms underpinning the size regulation *per se*.

One way to analyze the coupling of growth and division is based on modeling cell populations. Herein, the cell population behavior is deduced from modeling many individual cells, where the single cell models differ slightly when compared to one another. This can be achieved through e.g. a stochastic component in the model or the individual models are implemented to be subversions of one another. All the individual models together are called an ensemble and accordingly, the approach is called ensemble modeling (Henson, 2003).

Cell growth is dependent on nutritional conditions and the cell's metabolic capacity. Published models that simultaneously study growth and cell division usually do not consider central metabolism. In contrast, metabolic models, based on genomic and biochemical knowledge, are used to study global metabolic phenomena, e.g. using flux balance analysis (Pfeiffer et al., 2001; Kauffman et al., 2003). However, these models do not take into account the costs and benefits of specific proteins and enzymes that contribute to metabolism, such as ribosomes. They also disregard the impact of cytosolic space. Yet, recently, a minimal model of basic metabolism of a self-replicating system (such as ribosomes) has been developed to study growth rate related metabolic effects in microbes (Molenaar et al., 2009).

To study the mechanisms of size regulation in budding yeast, we employ a similar basic metabolic circuit where growth is an emergent property of the system itself. However, successful modeling of cell growth and size regulation must account for all three components mentioned above. Thus, we construct a model of the cell division cycle combined with a central metabolism component and use ensemble modeling to study population behavior (Spiesser et al., *in preparation*).

Concerning DNA replication, only recently models emerged that focus on the events occurring in the S phase of the cell cycle. Barberis and Klipp (2007) developed a coarse-grained, probabilistic model, simulating the difference in origin activation efficiency in budding yeast cells grown in glucose and ethanol. Furthermore, there exists a spatiotemporal model for DNA replication in mammalian cells (Takahashi, 1987) and some models of varying detail for *Xenopus laevis* (Bechhoefer and Marshall, 2007; Yang and Bechhoefer, 2008; Goldar et al., 2008). We constructed a fine-grained model for the spatiotemporal organization of DNA replication, providing the first means for detailed systemic analysis of this process in budding yeast (Spiesser et al., 2009). Later on, others became available (Brümmer et al., 2010; Yang et al., 2010; de Moura et al., 2010). No model for a detailed description of the elongation process during DNA replication could be found in the literature. Thus, our stochastic model for the replication machinery motion (elongation) remains the only one currently available (Spiesser et al., 2010). The models of DNA replication and elongation could, in the future, be combined with existing cell cycle models, e.g. Chen et al. (2004), to form a more accurate description of the cell cycle and to provide a more comprehensive insight into the crucial process of DNA replication.

Also, although other association studies between replication initiation sites and different genomic aspects, such as nucleosome positioning, have been conducted (Berbenetz

et al., 2010), we are the first to describe the functional relationship of genes that physically associate with replication start sites (Spiesser and Klipp, 2010). This thesis presents the models for size regulation (1), DNA replication organization (2) and elongation (3), as well as the functional gene-origin association study (4) in sequential order.

1.1.3 Organization of the Thesis

Systems biology is an interdisciplinary scientific field that combines theoretical and experimental research to provide an integrative view of complex biological systems. It requires an extensive knowledge of the biology, the modeling theory as well as the mathematical methodology. In order to provide the necessary background information, the three topics are introduced in chapter 1, sections 1.2, 1.3 and 1.4. Section 1.2 contains an introduction to the biological field of size regulation and DNA replication in budding yeast. The focus lies on introducing unresolved questions and biological issues regarding both systems. Here, the reader is also provided with detailed background knowledge regarding size homeostasis in yeast cell populations, the spatiotemporal organization of DNA replication, details concerning replication elongation and genomic characteristics of DNA replication initiation sites. This background knowledge is relevant for later chapters. Section 1.3 outlines the research field of systems biology and provides information on modeling theory. Section 1.4 concludes the introductory part, holding information about modeling with ordinary differential equations, statistical and basic stochastic concepts and the process of model parametrization.

After this introductory chapter, four chapters follow that are grouped according to the main research projects (1-4). Project (1), presented in chapter 2, concerns the study of size regulation and homeostasis of yeast cell populations. The chapter contains information about the model implementation and modeling assumptions as well as about the strategy to deduce complex population behavior from a single cell model. The parameter fitting and the model validation are also presented here. Subsections of chapter 2 show that the model suffices to reproduce characteristic aspects of the cell cycle and that size regulation on the single cell level is not needed for population size regulation. Furthermore, evidence it presented that the model qualitatively predicts the effect of altered growth conditions without nutrient sensing, suggesting that G_1 and G_2 regulation takes place simultaneously and that the average cell size converges to a point attractor.

Chapter 3 outlines project (2). It gives details on spatiotemporal modeling of DNA replication with focus on model implementation as well as available data for model validation. Furthermore, model simulations and experimental data for wild type and mutant conditions are presented. The third chapter includes subsections covering the impact of origin deletion on DNA replication and shows simulations of a systematic loss of origin function.

Chapters 4 and 5 are dedicated to research projects (3) and (4), respectively. They highlight details concerning the modeling of the replication machinery motion and a functional analysis of the gene content that is physically associated with replication initiation sites. Chapter 4 introduces the model and the modeling assumptions and describes the model fitting and ranking procedure. In addition, evidence is presented

that during DNA replication the elongation times are directly related to the segment lengths for most of the genome and informs the reader about how regions with strongly altered elongation distinctly map onto the budding yeast genome. Chapter 5 outlines the functional analysis of origin related genes (project (4)).

In the sixth and last chapter of this thesis the combined research approach and the main results are discussed. The chapter also informs the reader about future developments and the contribution of the presented work to the field of size regulation and DNA replication in a cell division cycle context. All supplementary information is provided in the appendix.

1.2 Biological Background and Research Projects

1.2.1 Cell Growth and the Cell Division Cycle

Coordination of biomass increase and cell proliferation is a fundamental process of life. It is characterized by a tight coupling between growth and the cell division cycle that ensures that only cells with sufficient nutritional supply in favorable condition commit to cell division (Alberts et al., 2007). While the mechanistic architecture of the cell division cycle's regulatory machinery has been studied intensely, a critical question remains in how the cell ensures this tight coupling between cell growth and cell division - hence maintaining their size constant over generations.

In budding yeast, cells divide asymmetrically (Hartwell and Unger, 1977). At START, i.e. the G_1 to S phase transition of the cell division cycle, the mother cell polarizes its actin cytoskeleton to grow a bud by polar secretion through a narrow bud neck, initiates DNA replication and spindle pole duplication, all of which occur during the S phase (Hartwell et al., 1974). After a G_2 phase of comparatively constant duration, during which the continued polar growth leads to a volume increase (primarily) in the bud, the cell enters the M phase and initiates rapid mitotic division, in which the bud receives a nucleus, separates from the mother and is born as a daughter cell. Due to the asymmetric division, the daughter cell is born at $\sim 60\%$ of the mother cell volume and hence, needs to acquire more mass and volume before it is ready to pass through a subsequent cell division cycle (Di Talia et al., 2007; Cookson et al., 2009). This is reflected in a prolonged duration of G_1, which is recognized as the primary phase for cell size regulation in *S. cerevisiae* (Brewer et al., 1984). The differential time spent in G_1 is required to maintain the size homeostasis within the population and it ensures that cells enter S phase only when having acquired sufficient resources to pass through it. Consistently with the G_1 phase being the primary window for size regulation, many size affecting mutations have been tied to ribosome biogenesis or a function in the regulatory network upstream of START (Jorgensen et al., 2002).

The regulatory network upstream of START has been intensely studied (reviewed in Bloom and Cross (2007)). The earliest undisputed activator of START is Cln3. Peak *CLN3* expression coincides with the mitotic exit and the transcriptional program that is part of resetting the cell for the next cell division cycle. Cln3 associates with the Cdk1 which is then able to phosphorylate the transcriptional repressor Whi5, thereby reliev-

ing the inhibitory effect on the heterodimeric transcriptional activators SBF and MBF (de Bruin et al., 2004). Their activation orchestrates the transcriptional components of START, leading to a peak in expression of a large number of genes (Spellman et al., 1998). Among their key targets are the remaining two G_1 cyclins: *CLN1* and *CLN2*. Like Cln3, Cln1 and Cln2 associate with Cdk1 and promote repression of Whi5. The result is a positive feedback loop that stabilizes once a critical threshold of Cln1/2 is reached (Skotheim et al., 2008). High Cdk1-Cln activity triggers polarization directly or indirectly. Furthermore, it triggers bud emergence, as well as spindle pool duplication and DNA replication (Enserink and Kolodner, 2010). The G_1 network provides a window in which the cells are responsive to mating pheromones, *via* the Cdk1-Cln inhibitor Far1 (Peter and Herskowitz, 1994). Similarly, it allows for size regulation, although the molecular mechanism that ties time and size together remains enigmatic. Several components such as Whi3 and the molecular chaperone Ydj1 have been proposed to act upstream of Cln3 and to integrate information on cellular size, although their roles remain unclear (Aldea et al., 2007). However, cell size distribution is known to vary with growth rate, leading to the suggestion that the ribosome biogenesis rate determines the critical size threshold, which would account for the fact that cell size is smaller at lower growth rates (Jorgensen et al., 2004). Hence, "size" regulation may depend on other factors than size *per se*.

The current understanding of size regulation builds on a critical size defined by the extensive regulatory network upstream of START (Barberis et al., 2007). However, it should be noted that all components upstream of the G_1 cyclins (Cln1 and Cln2) are dispensable for both viability and size regulation (Jorgensen et al., 2002; Enserink and Kolodner, 2010). While several of these deletions lead to altered average sizes and increased variation in size, size regulation *per se* appears to be retained. Thus, despite the combined efforts of the community, the nature of the core sizer remains obscure. This has led us to explore the possibility that size regulation occurs on a more fundamental level, using a minimal core model of cell growth, metabolism and division.

In the work outlined in chapter 2, we have created a minimal model to validate the hypothesis that the early cell cycle regulatory network is essential for size regulation. Surprisingly, our results clearly falsify this generally accepted hypothesis as size regulation on the population level occurs even when cells lack any means to determine their sizes. It is noteworthy that a very coarse grained model, which only allows for the allocation of resources in two biomass pools, surface area and metabolic capacity, suffices for population level size regulation. Importantly, the model has only eight parameters, lacks regulatory circuitry, is stable over a wide range of growth rates and robust against perturbations, after which the population quickly converges back to a stable attractor characteristic for its growth rate. Moreover, it accurately describes a number of observations on the cellular level, such as (i) the difference in G_1 duration between mother and daughter cells, (ii) the size increase with genealogical age in the mother line and (iii) the increase in size with increasing growth rate. However, we also fail to validate the hypothesis that G_1 regulation itself suffices, as an increase of growth rates requires an adjustment of the G_2 length to maintain the substantial difference between mothers and daughters in G_1. While inconclusive, this finding suggests that size regulation *in vivo*

occurs both in G_1 and G_2 also in *S. cerevisiae*, which is consistent with the observed increase in G_2 length in slowly growing cells (Barford and Hall, 1976). In conclusion, we show that size regulation in populations does not require any ability to sense size and our results strongly suggest that size regulation is an emergent property of growing and dividing cells.

1.2.2 Timing DNA Replication in Budding Yeast

DNA replication is one of the critical processes during the cell cycle progression that, if deregulated, can lead to checkpoint activation and cell cycle arrest (Alberts et al., 2007). The genomic duplication requires a complex coordination of successive events to initiate DNA replication and to distribute fully replicated chromosomes into the daughter cells (Bell and Dutta, 2002; Diffley and Labib, 2002). The initiation of DNA replication temporally stretches from the M phase over the G_1 phase into the early S phase. However, the chromosomal duplication is confined to the S phase of the cell cycle. Successful replication requires that the entire genome of an organism is duplicated without errors in a timely fashion only once per cell cycle. Therefore, DNA replication has evolved into a tightly regulated process, involving the coordinated action of numerous factors.

In prokaryotes, replication starts from a single well-defined site and proceeds with a speed of up to 500 nucleotides per minute until it terminates at the end of the genome. This mechanism leads to a homogeneous replication pattern that is identical in every cell cycle. The genome of the eukaryotic *S. cerevisiae* consists of 16 chromosomes, spanning a total length of about 13.5 million base pairs (bp) and if the replication machinery was to use the same single site strategy, DNA replication would take several days to complete. On account of this, replication of eukaryotic genomes initiates from multiple discrete sites scattered across the chromosomes, so called origins of replication (Stinchcomb et al., 1979; Zannis-Hadjopoulos and Price, 1998; Françon et al., 1999).

During the G_1 phase of the cell cycle, replication origins are prepared to fire, a process that is referred to as origin licensing (Weinreich et al., 2004). Although nearly all origins are licensed, only a selection of them is eventually destined to fire (Shirahige et al., 1993). Origin firing is also called initiation or activation and the ensemble of activated origins is the S phase specific firing pattern. In case an origin fires, two replication forks emerge from the origin, traveling along the DNA in opposite directions. The replication process continues until the whole DNA is replicated (Fig. 1.2). It becomes apparent that the firing pattern and with it the density of active replication origins in the chromosomes of eukaryotic cells determines S phase dynamics (Bielinsky, 2003). Accordingly, a direct correlation between the length of the S phase and the number of activated origins has been demonstrated in *S. cerevisiae* (van Brabant et al., 2001). Furthermore, it has been shown that there is a hierarchy of preferential origin firing that correlates with local transcription patterns (Donato et al., 2006) and that chromatin structure modulates origin activity (Tabancay and Others, 2006). Yet, it is still not known, how exactly origins are selected and how differential selection patterns shape S phase dynamics.

Experimental and computational studies have identified and mapped over 700 potential origin function target sites on the genome of *S. cerevisiae* (Feng et al., 2006;

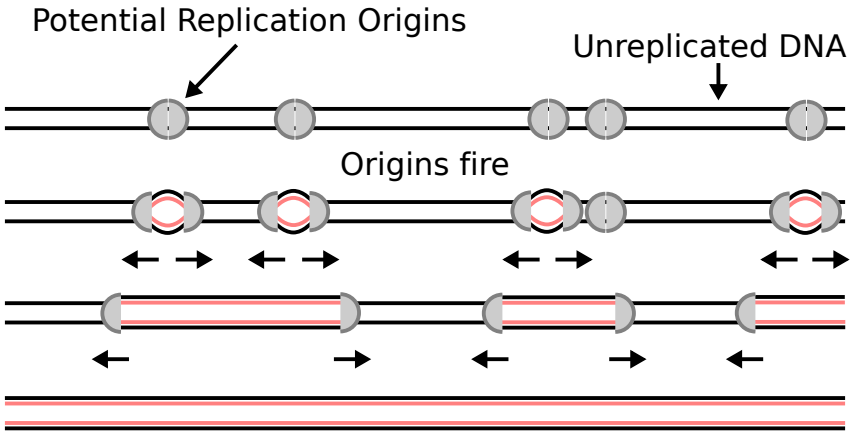


Figure 1.2: **Scheme of DNA replication process.** A subset of potential origins of replication fires at some time in S phase. Two replication forks issue from the activated origin, replicating the DNA in opposite directions at a certain rate along the DNA template stand. When two replication forks meet, they merge. The process continues until the whole DNA is replicated.

Nieduszynski et al., 2006; Raghuraman et al., 2001; Wyrick et al., 2001; Xu et al., 2006; Yabuki and Terashima, 2002). A number of studies have suggested that yeast chromosomes contain early and late replicating domains and exhibit replication timing profiles that are consistent with a highly regulated chronological program (McCune et al., 2008; Nieduszynski et al., 2006; Yabuki and Terashima, 2002), which is reproducible even under altered conditions (Alvino et al., 2007). Such nearly homogeneous replication kinetics favor the argument that, in budding yeast, the origins of replication fire according to a deterministic temporal program, as it has been reported for bacterial replication (Jacob and Brenner, 1963). Yet, considering the population averaged nature of the timing studies, there is a probabilistic quality in their replication profiles, suggesting that there might be variability among the replication programs of different cells. Indeed, recent studies have revealed an intrinsic temporal disorder in the replication of yeast chromosome VI, suggesting that there is no obligate order of origin firing (Czajkowsky et al., 2008). Under this premise, DNA replication appears to be essentially probabilistic in individual cells, instead of following a specifically regulated program. Nonetheless, origin firing patterns exhibit temporal tendencies over extended domains in cells that are at the same stage of replication progression (Czajkowsky et al., 2008), indicating the existence of similar spatial trends in budding yeast. A strong stochastic influence is indeed part of the replication program of its distant cousin fission yeast (Patel et al., 2006). Therefore, the observation of a stochastic component in the replication program would place budding yeast yet closer to the other eukaryotes, where it has been considered to be rather the exception in the general organization of eukaryotic replication (Rhind, 2006). In summary, even though intensively studied, the spatiotemporal organization of

the selective origin activation in *S. cerevisiae* remains unclear.

In chapter 3, I present a deterministic model for the DNA replication process in *S. cerevisiae*. It allows us to study the impact of variations in the temporal sequence of origin activation on DNA replication dynamics (Spiesser et al., 2009). The model is based on replication parameters that characterize every origin within the yeast genome: the position in the genome, the activation time of the origin and the emanating fork rate (elongation rate). Chromosomal positions and firing times for a certain number of origins have been reported (Nieduszynski et al., 2007) and fork rate values are available (Rivin and Fangman, 1980; Raghuraman et al., 2001; Yabuki and Terashima, 2002). Another parameter influencing the replication process, the origin efficiency, is not included in the model as an adjustable parameter, but implicitly incorporated. This is because only few data are available about individual origin efficiencies, which refer to the frequency at which an origin initiates DNA replication within a population of cells (Yamashita et al., 1997). The model is able to reproduce experimental data in the form of replication profiles in wild type and mutant conditions. We monitor the dynamics of the chromosomal duplication during simulations of wild type and perturbed replication conditions to assess the impact of differential origin activation patterns. Furthermore, we perform simulations of systematic origin deletion in order to provide predictions, which could be tested experimentally. This work aims at exploring the organization of the DNA replication program in budding yeast.

1.2.3 Elongation: DNA Replication Machinery Motion

The formation of the new DNA strands is a process called elongation. A central role in this process is played by activation of helicases, which break the hydrogen bonds holding the two DNA strands together and generate two single strands of DNA. In budding yeast, the origin recognition complex (ORC) recognizes the replication origin and then initiates the Mcm2-7 helicase loading in G_1 phase by recruiting specific licensing factors to form the pre-replicative complex (Dutta and Bell, 1997; Bell and Dutta, 2002; Stillman, 2005). When cells enter S phase, components of the pre-replicative complex are phosphorylated by kinase complexes: Cdk1-Clb5/6 and Cdc7-Dbf4 (Aparicio et al., 1999; Zou and Stillman, 2000). The phosphorylation regulates the Mcm2-7 helicase activity (Nguyen et al., 2000; Francis et al., 2009). Once activated, Mcm2-7 unwinds origin DNA to trigger the initiation of DNA replication (Weinreich et al., 2004; Takeda and Dutta, 2005).

The unwinding of DNA at the origin and synthesis of new strands form a replication fork at which the replication takes places in a non-symmetric manner. In the $5' \rightarrow 3'$ direction, the new DNA strand, also called the leading strand, is synthesized in a continuous manner by the DNA polymerase ϵ (Nick-McElhinny et al., 2008). In contrast, the DNA strand at the opposite side of the replication fork, the lagging strand, is formed in the $3' \rightarrow 5'$ direction. Because DNA polymerase ϵ cannot synthesize in this direction, DNA along the lagging strand is synthesized in short segments known as Okazaki fragments (Okazaki et al., 1967; Ogawa and Okazaki, 1980). In this process, the DNA polymerase α-primase complex builds RNA primers in short bursts along the lagging

strand, enabling the DNA polymerase δ to synthesize DNA starting from these primers in the $5' \rightarrow 3'$ direction (Nick-McElhinny et al., 2008). Afterwards, the RNA fragments are removed and the DNA ligase joins the deoxyribonucleotides together, completing the synthesis of the lagging strand (see Kunkel and Burgers (2008); Burgers (2009) for recent reviews).

In general, two replication forks emerge from an activated origin of replication, moving in opposite directions, as shown in Figure 1.2. The elongation rate (the rate at which the DNA is replicated) can differ between replication forks issued from the same origin, as well as for those from the other origins located on the chromosome. This results in a broad distribution of replication fork rates in budding yeast (Fig. 1.3).

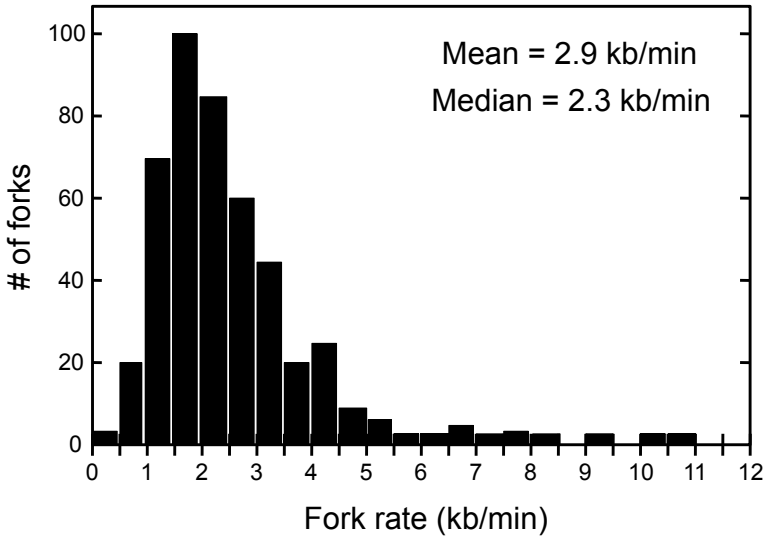


Figure 1.3: **Distribution of replication fork rates in kilo bases (kb) per minute (min).** Mean and Median of the distribution are indicated as well. Figure taken from Raghuraman et al. (2001).

The different fork rates at different chromosome regions could have either regulatory functions or could be caused by higher order structures of the chromosome (e.g. protein binding or tertiary structure). It has been suggested that epigenetic alterations influence fork rates both in yeast (Wintersberger, 2000; Zhang et al., 2000; Ji et al., 2001; Mechali, 2001; Pasero et al., 2002; Antequera, 2004) and in higher eukaryotes (Zhou et al., 2005; Hamlin et al., 2008). Furthermore, it is known that transcriptional activity regulates the replication origin activity (Kohzaki et al., 1999; Nieduszynski et al., 2005; Mori and Shirahige, 2007) and possibly it also plays a role in altering the replication fork progression (Lucchini and Sogo, 1994; Deshpande and Newlon, 1996; Wellinger et al., 2006). Though, it is not clear whether it enhances the fork rate due to already partly unwound DNA or impeding it because the DNA is blocked by proteins involved in transcription. Thus, it has not yet been established satisfactorily, how or where exactly deviations in the replication fork rates occur.

Fork rates are generally established by a directed movement of the replication machinery along the DNA template. The polymerase has to advance nucleotide per nucleotide, making the process itself non-continuous. This stepwise character is due to the movement of the complex from a replicated nucleotide to the next unreplicated one (movement step), which is interrupted by the catalyzing activity, during which the complex is stationary on the DNA. During the stationary state, the replication machinery incorporates a nucleotide into the nascent DNA strand that corresponds to the one of the template. This process is subject to various fluctuations, like nucleotide-specific polymerization kinetics, substrate availability (diffusion of the nucleotides), mismatch control (wrong nucleotides arriving at the polymerization sites but not being processed) and malfunctions that potentially cause delays. This makes DNA replication motion at least partly a stochastic process that is dependent on sequence properties such as length and base composition. However, to which extend this contributes to the overall replication rate, remains unclear. Whether these sequence-specific attributes play an active role in the variation of DNA replication rates has, to our knowledge, not been investigated.

This has led us to build a stochastic model for the DNA replication motion in budding yeast (Spiesser et al., 2010). In the model, we interpret the replication machinery movement as a directed random walk. A directed random walk can be seen as a process in which the location of an object randomly changes by a single directed step, depending on a number of probability parameters. In the case of the replication machinery, the directed step is the movement with probability p or the stalling/waiting with probability $1 - p$, as depicted in Figure 1.4.

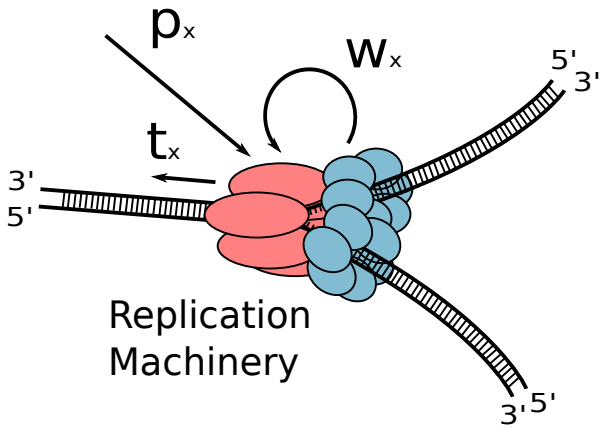


Figure 1.4: **Schematic view of the DNA replication machinery.** The replication machinery can move forward with a base-dependent probability p_X, taking a mean time t_X for the forward step and a mean time w_X for the waiting step for base $X \in \{A, T, G, C\}$.

The replication machinery only moves if the appropriate nucleotide is instantly available, can be incorporated without problems and stalls in case of a mismatch or other

fluctuations, as mentioned above. The movement of the machinery takes the characteristic time t and the stalling takes the time w. Probabilities (p), transition times (t) and waiting times (w) may be specific for the four bases A,T,G and C. We use the model to study the variation of DNA replication rates and in chapter 4, I present a concise characterization of sequence-specific replication rates, as well as a spatial map of regions with sequence-independent alterations in replication rates within the genomic landscape of budding yeast.

1.2.4 DNA Replication in a Genomic Context

In budding yeast, all replication origins share a common feature, an approximately 200 base pair sequence called autonomously replicating sequence (ARS) (Newlon and Theis, 1993). Within this region, an eleven base pair sequence, the so-called ARS consensus sequence (ACS) is specifically recognized by the ORC (Theis and Newlon, 1997). A sequence match to the ACS is essential, although the presence of this element alone does not define origin function *per se* (Breier et al., 2004; Nieduszynski et al., 2006). Furthermore, yeast origins consist of three elements that, while non-essential, contribute to origin function (Marahrens and Stillman, 1992). Thus, origin function is an evolutionary conserved sequence feature and it seems possible that also sequences in origin vicinity show some functional conservation.

Generally, for origin initiation, Cdk1 activity requires binding of one of the cyclins Clb6 or Clb5. It is known that Cdk1 is active throughout S phase, however the corresponding cyclins change. As schematically shown in Figure 1.5, Clb6 is expressed and bound to the kinase in the first half of S phase to ensure its activity. Clb6 gets degraded near mid S phase and the cyclin Clb5 binds Cdk1 (Jackson et al., 2006). Both complexes (Cdk1-Clb6 and Cdk1-Clb5) can activate replication origins (Epstein and Cross, 1992; Kühne and Linder, 1993; Schwob and Nasmyth, 1993). Due to constant Cdk1 activity, origins initiate DNA replication throughout the entire S phase of the cell cycle. Although, most origins fire near mid S phase, it has been argued that there are chromosomal regions that can be classified into early and late replicating domains (Yabuki and Terashima, 2002; Nieduszynski et al., 2006; McCune et al., 2008). Early origins initiate the replication in the first half of the S phase (early domains) and late origins in the second half (late domains). Correspondingly, genes that are located close to origins are duplicated early or late as well. McCune and colleagues have studied DNA replication in a *clb5*$\triangle$ environment and thus, altered Cdk1 activity in the second half of the S phase. They have demonstrated that only for a defined subset of origins the initiation time is altered in this condition (McCune et al., 2008). They labeled regions in the genome that showed altered replication kinetics in the *clb5*$\triangle$ mutant as Clb5-dependent-regions (CDRs) and those unaffected as non-Clb5-dependent-regions (non-CDRs). For further details see section 3.3.2.

As mentioned in section 1.2.2, origins initiate DNA replication at a fixed time during S phase (Raghuraman and Brewer, 2010). However, what exactly determines this fixed time is not known. Rhind et al. (2010) made a step in the direction of understanding origin timing by specifying origin firing times as intrinsic relative firing probabilities. The

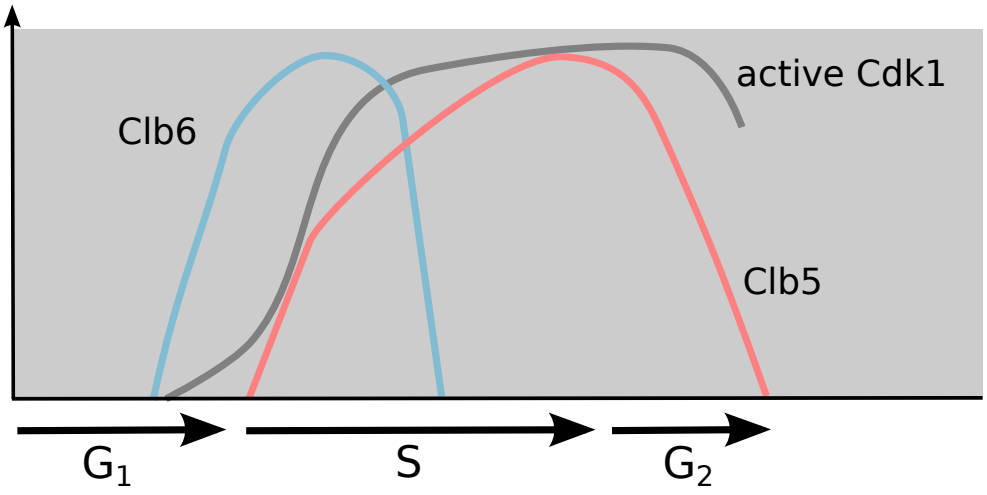


Figure 1.5: **Schematic view of Cdk1 activity in association with the cyclins Clb5 and Clb6.** Clb6 binds the kinase in the first half of the S phase but is degraded near mid S phase. In the latter half of the S phase the cyclin Clb5 associates with the kinase, ensuring its constitutive activity. Thus, Cdk1 is active throughout the entirety of S phase.

origins with a relatively higher probability are more likely to fire early in S phase while origins with a relatively lower probability are rather unlikely to do so. However, the regulatory event which eventually determines the relative probabilities and with it the timing of origin initiation, still remains obscure. Several mechanisms have been proposed that could potentially account for variations in the relative origin firing probability. Nucleosome positioning, chromatin status, transcription and the number of Mcm2-7 molecules loaded onto the DNA are amongst them (Berbenetz et al., 2010; Rhind et al., 2010). It has been observed that heterochromatin replicates late (Goren and Cedar, 2003), which is consistent with a view in which chromatin density delimits the accessibility of replication origins. Indeed, in budding yeast the chromosomal context influences the origin firing time (Ferguson and Fangman, 1992; Friedman et al., 1996). Consistently, a correlation between larger, transcriptionally active regions and early replication has been observed in *Drosophila* (MacAlpine et al., 2004), supporting the idea that an open chromatin structure facilitates origin activation and thus, earlier firing. Altogether, the time of replication initiation is potentially governed by a combination of factors that act within the genomic origin domain. They influence the firing probability and thus, the activation time might be mirrored to some extent in properties of the imminent origin vicinity. Functional genomic analysis could shed some light onto the nature of this influence. Furthermore, it might help to understand how origin sequences have evolved and with it the effects on the replication program (Raghuraman and Brewer, 2010).

An interesting aspect of genomic duplication is that genes in close origin proximity are

replicated first. The duration of S phase in budding yeast is approximately 45 minutes (Barford and Hall, 1976). This means that genes that are duplicated in the first minutes of S phase are present in two copies for a much longer time than the genes that replicate late. It has been shown that the copy number of genes can have a great influence on cellular behavior, a so called gene dosage effect (Di Talia et al., 2007). Therefore, it seem possible that there is a functional relationship of genes that are close to replication origins due to (i) functional conservation of origin sequences and potentially also sequences in their vicinity and (ii) evolutionary optimization for positive gene dosage effects during S phase. So far we can only speculate about it because the functional relationship of genes in origin proximity has not been investigated yet. However, if gene function is conserved at all, then there might also be a difference between genes in early (non-CDRs) and late (CDRs) replicating domains. Furthermore, the idea is interesting from a reverse engineering point of view: could it be possible to predict whether an origin initiates replication early or late in S phase on the basis of the genes in its vicinity? These considerations have led us to investigate the gene environment of origins. To this purpose, in chapter 5, I present an analysis for the functional relationship of genes that co-localize with origins of replication. We analyze the gene function in origin proximity using a gene ontology term enrichment test. Furthermore, we analyze the genes that are localized in CDRs and non-CDRs.

1.3 Methodological Background

1.3.1 Systems Biology

Systems biology is a field of research that is driven by the aim to understand the biochemical world of life and its generally applicable principles. Biochemical life, as we know it, is highly diverse, and compared to other physical systems, it seemed for a while to exist on the borderline of chaos. On first sight, every species, even every organism seemed different to the next one and biochemical bonds that constitute the basis in this construction kit lead to even more diverse and complex molecules with seemingly uncountable functionalities and properties.

Life is but a small subset of all chemical and physical systems and with its inherent diversity it seemed not to be the most attractive one for the search of fundamental principles. However, biology has its laws, underlying principles and generalities as well. They began to arise in the nineteenth century with studies from Lamarck and Darwin, who provided evidence that species had common origins and thus, were much more similar than originally anticipated (Lamarck, 1809; Darwin, 1860). Fundamental principles in biology were discovered. For example, a central dogma emerged on the molecular level of life, i.e. that DNA encodes for proteins and that the information encoded within is transmitted *via* mRNAs (Crick, 1970).

Cell and molecular biologists engaged in entangling the chaotic wirings of cellular networks in the strive to understand the processes that constitute life. They did so by identifying the single components and their interactions, assuming that the nature of complex things is to be understood by reducing them to the interactions of their parts

(reductionism). In contrast, systems biology antagonizes this point of view, assuming that essence of complex systems can only be captured by looking at the system as a whole (wholism), or as Noble (2008) defines it:

> "Systems biology [...] is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programs, but different [...]. It means changing our philosophy, in the full sense of the term".

Thus, systems biology is a novel paradigm that extends biological research aiming to uncover fundamental principles and to reveal emergent properties of complex interacting systems by relating systemic properties to interactive properties of the single components (Westerhoff et al., 2009). The concept of a systemic approach refers to an ancient philosophical point of view, which was argued for more than 2000 years ago in a treatise named *Metaphysics*, formulated by Aristotle and concisely summarized by:

> "The whole is something over and above its parts, and not just the sum of them all." (Aristotle, Book H, 1045:8-10 in Jaeger (1957)).

I argue that systems biology is the modern implementation of this ancient philosophical point of view, that in present times could emerge due to the availability of new and more powerful tools for systemic research and data generation.

1.3.2 Modeling in Biology

The key concept of systems biology is mathematical modeling. It is a powerful tool that uses mathematical language for the description of biological phenomena. Herein, the model represents the current knowledge of the biological system in an abstract, usable form. Thus, mathematical models allow for formal descriptions of hypothesis and their rigorous testing by comparison of model simulations with data from various experimental sources. Integrating a computational approach and experimental research is crucial to understanding complex biological networks (Kitano, 2002).

In the beginning there is usually a hypothesis that arises from a question or contradictory issue about a biological system. The biological system itself can be seen as the center of an imaginary scientific workflow (such an idealized systems biology workflow is schemed in Figure 1.6) in which model, data and hypothesis are refined in an iterative cyclic process with the final aim to gain knowledge about said system. Note that in reality, scientific research is neither strictly cyclic nor straightforward (Alon, 2009). In detail, the working hypothesis is formalized on the basis of the current biological knowledge by using an appropriate modeling framework and tested by comparison to experimental data. In the iterative revision process, the constructed model is tested and refined until it satisfactorily reproduces the experimental evidence. Ideally, the process leads to a general refinement of the hypothesis and the model and to the generation of new experimental data. The new experimental data should, in that case, be designed to validate or reject model predictions.

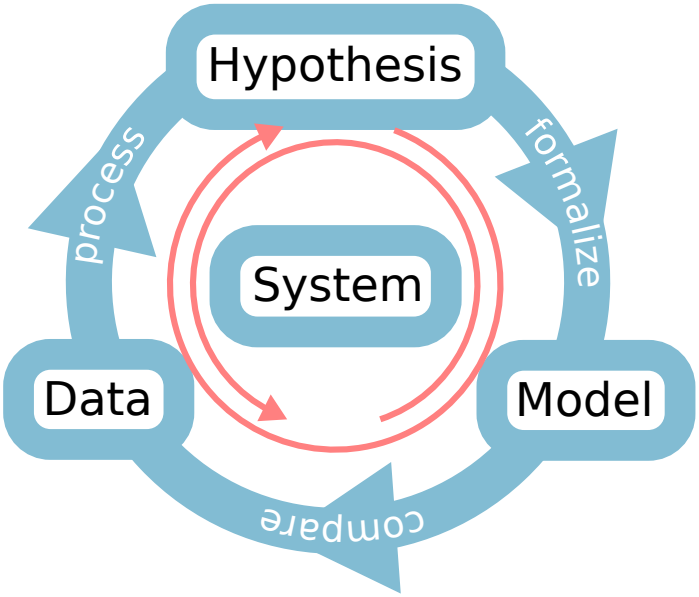


Figure 1.6: **Idealized workflow of a systems biology research approach.** In an iterative cycle of model construction/refinement and model - data comparison, hypothesis are tested and the knowledge about a specific system is extended.

At this point, I would like to draw the attention to an important aspect in the systems biology workflow. The stage of hypotheses formalization, i.e. the model construction. There are various ways of model construction. By using an appropriate formalism nearly all systems imaginable can be modeled. However, it is also important to note that modeling a desired system can be achieved in more than one way, meaning that the same system can be modeled with different approaches, highlighting different systemic properties. Thus, choosing the appropriate formalism is a crucial step. Every formalism has its advantages and disadvantages and it is important to balance model granularity (detail) and model complexity (manageability).

A commonly used formalism is the modeling with ordinary differential equations (ODEs, see section 1.4.1). ODE models allow for the dynamic, time-dependent continuous representation of biological systems (Klipp et al., 2005). Also dynamic, but more complex is the representation with partial differential equations (PDEs), which describe the dependencies of modeled entities on time and on space (Hjortso and Bailey, 1983). Both, ODE and PDE models are advisable when dynamic properties of systems are under investigation. However, it should be noted that both rely heavily on the availability of experimental data for adequate parametrization, which is sometimes spares and hence, can be a drawback (see section 1.4.3). It should also be mentioned that modeling with ODEs assumes that molecular quantities in the systems are relatively high and fluctuations that might have an influence at low abundance, can be

neglected. Furthermore, there are Boolean Networks, a formalism for discrete modeling that is used when modeling entities that can only be either on or off (0 or 1). For example, this is often assumed to be the case in gene regulatory networks (Li et al., 2004). Boolean Networks do not need parametrization and give information on properties of the network structure. Dynamic features of a network can also be explored with Boolean Networks. Petri nets are a formalism for a partly-continuous representation of systems and thus, bridge the gap between continuous and discrete models (Sackmann et al., 2006; Mura and Csikász-Nagy, 2008). There are many other model types, amongst them models of delay differential equations (Boczko et al., 2005), statistical models (Zhao et al., 2001), stochastic descriptions of biological processes (Zhang et al., 2006) and constrained-based models (Rokhlenko et al., 2007) to name a few. For further reading on different mathematical biology and bioinformatics formalisms, please, refer to the comprehensive literature review from de Jong (2002) or to Szallasi et al. (2006).

Of particular interest for my work are ODE and stochastic models, two formalisms which I introduce in more detail in sections 1.4.1 and 1.4.2, respectively with a focus on those aspects relevant to chapters 2 and 4. In summary, the type of modeling framework depends on the type of biological system in question, its desired description and the problem that is to be studied. There is no such thing as only one correct way to do good research or to answer a scientific question and in essence, according to Box and Draper (1987):

> "all models are wrong; the practical question is how wrong they have to be to not be useful".

1.4 Mathematical Background

A particular challenge in systems biology is to understand how the interplay of single components and their interactions influence the behavior of a biological system in time and space. Modeling this can be achieved e.g. with differential equations. Differential equations describe the change of state variables that are dependent on values, such as time and space, which are the independent variables. Although other potential independent variables such as volume or temperature exist, they are often assumed to be constant or expressed as to be time and space dependent. If only changes dependent on one independent variable are to be tracked, one uses ODEs, whereas tracking of changes in time and space would require the use of PDEs. The following brief introduction which is mainly based on Szallasi et al. (2006) and Klipp et al. (2009), is focused on modeling with ODEs.

1.4.1 Modeling with Ordinary Differential Equations

Using ODEs, concentrations of substances are modeled as time-dependent variables. A variable concentration x at time t is determined by its initial concentration and a

differential equation of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = synthesis - degradation - phosphorylation + binding, etc. \tag{1.1}$$

Every single rate of a reaction (e.g. *synthesis, degradation*) or reaction velocity (v) may be expressed as a function of the compound concentrations $v = f(x, t, p)$, represented by a rate law, which may be time-dependent and have one or more rate constants (p).

The fate of a molecular interaction network with n species and m reactions can now be determined by a set of ordinary differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_1, ..., x_n, p_1, ..., p_j, t), \tag{1.2}$$

where $i = 1, 2, ..., n$ and x_i is the set of all variables, e.g. compound concentrations and p_j is the set of all parameters (rate constants) necessary to describe the reactions. Defining the column vectors $\mathbf{x} = (x_1, ..., x_n)^T, \mathbf{f} = (f_1, ..., f_n)^T$ and $\mathbf{p} = (p_1, ..., p_j)^T$, equation 1.2 reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{p}, t). \tag{1.3}$$

Furthermore, a comprehensive way to describe a biological network is *via* its stoichiometric matrix $\mathbf{N}$ ($n \times m$). Herein, a stoichiometric coefficient $N_{i,j}$ denotes the molecularity of compound i produced or consumed in a particular reaction j (Ingalls and Sauro, 2003). Consequently, the ODE system is expressed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{N}\mathbf{v} \tag{1.4}$$

with the column vector $\mathbf{v} = (v_1, v_2, ..., v_m)$, comprising all reaction velocities.

As mentioned above, the velocities are represented by rate laws. The rate laws are usually synthesis and degradation terms (forward and backward reaction) and hence, determine the speed of the modeled reaction. However, a rate law can further include other aspects of the system, e.g. specific reaction kinetics. Different reaction kinetics are used to describe different biological phenomena. E.g. a Michaelis-Menten kinetic is used to describe the reversible formation of an enzyme-substrate complex from a free enzyme and a respective substrate and an either reversible or irreversible release of a product from the enzyme (Michaelis and Menten, 1913). Standardized kinetics can be adapted and modified to be applied to specific reactions or various forms of allosteric regulations. All biochemical kinetics however are based on the mass action law (Waage P.; Gulberg, 1986), which states that the reaction rate is proportional to the probability of a collision of the reactants. This probability is in turn proportional to the concentration of reactants to the power of the number in which they enter the specific reaction. The mass action law for an irreversible reaction is given by

$$v = k_j \prod_{i=1}^{m_i} x_i^{s_i} \tag{1.5}$$

where $x_1, ..., x_{m_i}$ are the substrates and $s_1, ..., s_{m_i}$ the numbers of molecules at which the substrates enter reaction j.

Since the reaction rate is proportional to the probability of a collision of the reactants, volume changes alter the probability of such a collision. If the volume decreases the probability increases and *vice versa*. Thus, if volume changes are envisioned in a model, they have to be accounted for in the case of compound concentrations with an additional dilution term or in the case of absolute amounts the reaction rates have to be scaled according to the volume changes. In zero-order and first-order kinetics, where a reaction occurs at a constant rate

$$v = v_o \tag{1.6}$$

or proportional to only one substrate

$$v = k_j x_1, \tag{1.7}$$

respectively, a change in the volume does not have an impact on the reaction rate. However, as soon as a minimum of two compounds are involved in a reaction (second-order kinetics or higher) the volume (V) influences the probability that molecules meet in the order of number of molecules involved in the reaction, leading to

$$v = k_j \prod_{i=1}^{m_i} (x_i^{s_i} \frac{1}{V^{s_i}}) V. \tag{1.8}$$

There are different ways to find solutions to an ODE system. At times, an analytical solution of an ODE system can be found if the f_i's in equation 1.2 are linear functions. However, generally it is not possible to find analytical solutions and thus, with a vector of all initial concentrations ($\mathbf{x_0} = (x_1(0), ..., x_n(0))$) and the vector of all rate constants ($\mathbf{p}$) a numerical solution for the ODE system is computed and the transient compound concentrations can be simulated. One way to do so is to start at $x_i(0)$, choose adequately small $\triangle t$'s and employ

$$x_i(t + \triangle t) = x_i(t) + f_i(x_1(t), x_2(t), ..., x_n(t)) \triangle t \tag{1.9}$$

to follow the time courses of the systems compounds. Note that there are also other ways to compute numerical solutions (Petzold, 1983). The system can furthermore be analyzed with regards to its different properties, e.g. its steady states and their stability or their sensitivity against parameter change. Throughout this thesis, the simulation of the state variables in ODE systems has been performed using the ODE solver LSODA (Petzold, 1983).

1.4.2 Statistical and Basic Stochastic Concepts

Many biological processes occur in a non-deterministic fashion. Modeling with ODEs assumes relatively large quantities over which an average behavior is approximated, with which local fluctuations are averaged out. However, e.g. in case of low amount quantities,

fluctuations might play a more pivotal role. Those processes might be best approximated with stochastic descriptions. There is not one appropriate way to do so, but the type of description is chosen to reflect the properties of the system. In the following, I outline the statistical and basic stochastic concepts that are relevant to later chapters of this thesis. The theoretical formulation is based on Abramowitz and Stegun (1972); Stirzaker (2005) and Klipp et al. (2009).

In probability theory, a random variable (X) is a variable whose value results from the outcome of some type of random experiment. The value of the random variable is not fixed but changes with the experiment performed. It is, so to say, one realization of a random process. The random process can be characterized by the probabilities for the random variable that different values occur. The range of probabilities which the random variable can adopt are described using a probability distribution. In case that the random variable is strictly real-valued, i.e. the sample space $\Omega = \Re$, which is the case in most practical applications, the probability distribution is completely described by the cumulative distribution function. Generally, the cumulative distribution function is used to express the probability to obtain a certain real-valued random variable X at a certain value x (or less than x) under a given distribution and has the following form

$$\mathbf{F}_X(x) = P(X \leq x). \tag{1.10}$$

The derivative of the distribution function ($\mathbf{F}'_X(x)$) is called the probability density function

$$\begin{aligned} \mathbf{F}'_X(x) &= [f_X(x)]_{-\infty}^{x} && (1.11)\\ &= f_X(x) - f_X(-\infty) && (1.12)\\ &= f_X(x), && (1.13) \end{aligned}$$

so that

$$\mathbf{F}_X(x) = P(X \leq x) \equiv \int_{-\infty}^{x} f_X(t)\mathrm{d}t. \tag{1.14}$$

If X is discrete, f_X is called the probability mass function of X and

$$f_X(x) = P(X = x). \tag{1.15}$$

The probability distributions are often characterized *via* their moments. I focus on the definition of the first two moments at this point. For further reading concerning moments and the moment generating functions see Stirzaker (2005) or Gardiner (2009). The first moment is called the expectation (or mean) and is defined as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x \equiv \mu. \tag{1.16}$$

The second moment is the variance, which is

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\mathrm{d}x. \tag{1.17}$$

The expected value gives an idea of the mean outcome of an experiment if all potential outcomes with their respective probabilities are taken together. In the finite sample case, the empirical mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.18}$$

where $x_1, x_2, ..., x_n$ are n outcomes or samples of a random experiment. As n increases, the empirical mean converges to the expected value, defined in 1.16, thus

$$\lim_{n\to\infty} \bar{X} = E(X). \tag{1.19}$$

The variance denotes the mean squared deviation of the outcomes from the mean outcome, i.e. the spread of the probability distribution from the mean. The sample variability, i.e. the distortion of every sample from the sample mean, for a finite amount of samples is often expressed as the empirical standard deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2}. \tag{1.20}$$

The standard deviation is the square root of the variance, thus $\sigma^2 = Var$. Another measure of the variability is the median absolute deviation,

$$MAD = median(|x_1 - x_{med}|, ..., |x_n - x_{med}|), \tag{1.21}$$

where the median (x_{med}) is defined as the value that is greater than or equal to 50% of the sample elements.

As mentioned above, there are discrete and continuous probability distributions. As an example of a discrete probability distribution I show here the binomial distribution. In an experiment with n independent trials, that have a probability p of success and a probability $1-p$ of failure in each of the trials, the random variable X of x successes is binomially distributed and has the form

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x},\ 0 \leq x \leq n. \tag{1.22}$$

A continuous and most frequently used distribution is the normal (Gaussian) distribution. Normal distributions find many applications in statistical testing in various fields, biology amongst them. Residuals in error distribution are often assumed to be normally distributed (see chapter 4) and in density estimations of unknown distributions, estimators often have kernels of normal distributions (see chapter 5). The reason for this is that, according to the central limit theorem, the sum of independently distributed random variables with finite means and variances will resemble a Gaussian distribution

(Laplace, 1812). The normal density for any constants μ and $\sigma^2 > 0$ is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \; -\infty < x < \infty. \tag{1.23}$$

Applications for the use of the binomial and Gaussian distributions can be found in chapter 4.

An important estimator for the probability distribution function of an unknown distribution is the empirical cumulative distribution function (ECDF). It has been shown that it converges towards the unknown underlying probability distribution (Cantelli, 1933; Glivenko, 1933) and has the following form

$$\mathbf{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{1.24}$$

where $I(A)$ is the indicator function (or characteristic function) of an event A.

Another important question concerns the relationship of multivariate samples. The measure that specifies the strength of such a relationship is called correlation. Throughout this thesis the Spearman rank correlation (Spearman, 1987) is used to measure correlations. The Spearman rank correlation is defined as

$$SC = \frac{\sum_{i=1}^{n} (r_i^x - \bar{r}^x)(r_i^y - \bar{r}^y)}{\sqrt{\sum_{i=1}^{n} (r_i^x - \bar{r}^x)^2 \sum_{i=1}^{n} (r_i^y - \bar{r}^y)^2}}, \tag{1.25}$$

where r_i^x,r_i^y are the ranks of x_i, y_i in the samples $x_1, x_2, .., x_n$ and $y_1, y_2, .., y_n$, respectively. $\bar{r}^x$ and $\bar{r}^y$ denote the mean ranks.

1.4.3 Model Parametrization

The parametrization of mathematical models is one of the major challenges in systems biology. Herein, it does not play a pivotal role what kind of model one seeks to adjust to the underlying biology. Whether one has to deal with an ODE model or with a stochastic one, whether Petri-nets are used or the model is formulated with PDEs, parameters need to be fine-tuned, so that a model reflects the major biological properties of a system and fits available data. Only then, the predictive power of a model can be assured and there might be a gain of knowledge through the iterative cycle of model prediction and hypothesis testing.

The parametrization of models is usually a non-trivial task, because very frequently biological data is sparse and a complete parametrization of e.g. an ODE only with already published kinetic constants seems almost impossible. Furthermore, a major pitfall in using published kinetic data is that even though constants might be available for a certain biological reaction, the constants might change with variations in temperature, pressure, pH-level, simply in a different experimental setup. Therefore, caution must be

taken in the usage of published parameters, since what might be the correct parameter for one modeling approach, might well be totally off the charts for a different one.

Parameters reflect certain properties of biological reactions, such as decay rates of proteins or mRNAs, diffusion rates of different compounds or rates that determine the velocities (v) of reactions. In a hypothetical, optimal experimental setup, parameters could be determined exactly and, given the correct formulation of a model, the experimental and the simulated response of a system would be exactly the same. However, in reality the data are noisy and contain errors. Those errors are generally assumed to be normally distributed random errors. Nonetheless, models are fitted to the data in order to approximate the real parameters, whether it is to adjust known parameters or to estimate unknown ones. The fitting is formulated as the following problem. Given m data points

$$(t_i, k_i),\ t_i, k_i \in \Re,\ i = 1, ..., m, \tag{1.26}$$

that describe the measurement of k_i at time points t_i and given that there is a dependency of k on t that might be expressed using a function f as

$$k(t) = f(t; \mathbf{p}) \tag{1.27}$$

where f depends on a set of n parameters $\mathbf{p}$, it can be possible (problem of parameter identifiability is discussed below) to find a set of parameters $\mathbf{p}$ such that

$$k_i = k(t_i) = f(t_i; \mathbf{p}),\ i = 1, ..., m. \tag{1.28}$$

This would the case if there were no measurement errors and if the model were formulated correctly. In reality however, we find that there are errors (ϵ), which is why, for $i = 1, ..., m$ it holds that

$$k_i = f(t_i; \mathbf{p}) + \epsilon_i \tag{1.29}$$

$$\epsilon_i = k_i - f(t_i; \mathbf{p}), \tag{1.30}$$

from which the following optimization criterion is deduced

$$\boldsymbol{\epsilon}^2 = \sum_{i=1}^{m} \epsilon_i^2 \stackrel{!}{=} \min. \tag{1.31}$$

According to the criterion the fit is done in a least-square fashion, i.e. finding a parameter set $\mathbf{p}$ that minimizes the sum of squared residuals (RSS) (Deuflhard and Hohmann, 2002). Assuming that the measurement errors follow a Gaussian distribution, that is to say that the noise has the same variance σ^2 for all the data points, the procedure is also called maximum likelihood method, where

$$\ln L(\mathbf{p}|\mathbf{k}) = -\frac{1}{2} \sum_{i}^{m} \left(\frac{k_i - f(t_i, \mathbf{p})}{\sigma} \right)^2 \tag{1.32}$$

is called the log-likelihood and the optimization is to maximize the likelihood function $L(\mathbf{p}|\mathbf{k})$ (Klipp et al., 2009). This formalism weights the error of each data point with its variance σ^2.

In case of linearity of the function f, classical approaches like the Gaussian algorithm can then be used to solve the system. However, even though the system might be linear, its solution f usually is not, which is why the solution often may only be found iteratively. The Gauss-Newton method ascribes the numerical solution of such a non-linear least-squares problem to the solution of a series of linear problems, which can be solved explicitly (Deuflhard and Hohmann, 2002). However, the Gauss-Newton method can only be applied to minimize the sum of squared functions. There are other approaches to tackle optimization problems. Generally, one differentiates between local and global optimization methods. Local optimizers, such as the quasi-Newton method limited-memory Broyden-Fletcher-Goldfarb-Shanno method for bound-constrained optimization (L-BFGS-B) (Byrd et al., 1995; Zhu et al., 1997), converge to the local extremum in the parameter space, which is closest to the provided starting point. Global optimizers scan the parameter space iteratively for a solution fulfilling the given optimization criterion and rank them according to an objective value. An example for a global optimizer is the probabilistic Simulated Annealing algorithm (Kirkpatrick et al., 1983). The fundamental difference between local and global optimization strategies is that the overall structure of the objective function is only known to the global methods, which is why they can evade local minima. A disadvantage of global methods is that they are, in contrast to most local methods, numerically expensive and only provide an approximation of the best solution to the problem. This is why a combination of the two approaches can be an advisable strategy. A global optimizer can be used to determine the global minimum of the parameter space, followed by a local optimization procedure, starting in the global minimum area to obtain the best possible result, the optimal fit (see chapter 4 for an application of this strategy). A different approach could be to sample random starting points in the parameter space and then to perform local optimizations. The results can finally be ranked by some criterion to determine the best fit. For further reading on optimization techniques and strategies, refer to Jarre and Stoer (2003); Moles et al. (2003).

A common pitfall in systems biology is the identifiability of parameters. In a case of non-identifiability, no unique solution to the optimization problem can be found, which implies that more than one parameter set fits the data equally well (see chapter 4). This phenomenon can have different reasons. The experimental data might be too few or the wrong kind of data, which would mean that the number of data points compared to the degrees of freedom of the model (parameters) might be too limited (overfitting). Or the given data does only suffice to parametrize a model with a different structure, for example a more coarse-grained model (overdetermined). That is to say that there might be data for e.g. the product u of a certain reaction $u = k_1 \cdot k_2$. This data might well allow for the adequate description of the product and the reaction itself, but it might not suffice to fully determine the parameters k_1 and k_2, for any combination of $k_1 \cdot k_2$ yielding u would be correct. In such a case, the parameters are said to show correlations. One simple solution to this problem is model reduction. In the example case, parameters

k_1 and k_2 could be replaced by a single parameter k_{12}, which could then potentially be fully determined. Model reduction might lead to the loss of detail when it comes to biological descriptions. However, it might also lead to an increase in predictive power of the model, since overfitting/-determination is avoided. Thus, during the process of model parametrization it so often happens that one has to compare models of different granularity or structure with one another. The assembly of competing models is called a model ensemble and information criteria, such as the Akaike Information Criterion (AIC) (Akaike, 1974) can be used for model selection. The AIC quantifies the information that is lost when an estimated statistical model is used to describe reality and combines this goodness of fit with the complexity (degrees of freedom) of the model. The model with the lowest AIC value of the model ensemble is the best. The AIC value is a relative measure and therefore, not suitable for single model evaluation but only ranking within a model ensemble. The AIC can be calculated on the basis of two different statistical measures, the RSS and the coefficient of determination (R^2) as follows

$$AIC = 2k + n\left[ln\left(\frac{RSS}{n}\right)\right] \tag{1.33}$$

with n equal to the number of observations and $RSS = \boldsymbol{\epsilon}^2$ as defined in equation 1.31. Furthermore,

$$AIC_{R^2} = 2k + ln\left(\frac{1-R^2}{n}\right) \tag{1.34}$$

with

$$R^2 = 1 - \frac{RSS}{\sum\limits_i (k_i - \overline{k})^2}$$

where $\overline{k}$ is equal to the mean of k, as defined in equation 1.18. For further reading on the topic of parametrization in modeling, optimization strategies and implications on model building, reduction and selection see Klipp et al. (2009).

2 Size Regulation is an Inherent Property of Budding Yeast Populations

In the following chapter, I present a modeling approach to explore the nature of size regulation in budding yeast. The model links the cell division cycle and central metabolism. Ensemble modeling is used to extrapolate population behavior from single cell properties. The chapter is based on:

T. W. Spiesser, M. Krantz and E. Klipp. Size regulation is an intrinsic property of growing cell populations and does not require any sensing or signaling events. *In preparation.*

2.1 Introduction

Maintaining an appropriate cell size is an integral component of every living cell. Generally, cell growth is regulated by a systemic linkage between the growth rate, the cell size and cell division (Ramanathan and Schreiber, 2007). The cell division cycle regulates all processes that are required for successful and timely cellular duplication and it is tightly coupled to cellular growth. Only if cells reach a certain size they are able to pass the START checkpoint and commit to cell division. How exactly cells coordinate growth and division however, remains enigmatic. Here, I present a mathematical model for cell growth to study the gating procedure that leads to the cell division commitment and, as such, the coupling of growth and division (see section 2.3.1).

The model, outlined in section 2.2, is based on ODEs complemented with a function that allows for transcription to occur in the form of random bursts. Cell population behavior is inferred from the single cell model through multiscale simulations, using a simulation environment, developed especially for this purpose (section 2.2.2). Simulations show that the model is in excellent agreement with cell division cycle and growth-related empirical data (sections 2.3.2 and 2.3.3), is stable over a wide range of growth rates (section 2.3.4) and robust against perturbations (section 2.3.5), despite the lack of a size sensing mechanism and size regulatory circuitry. The cell populations show size homeostasis, where the average cell size converges to a stable attractor characteristic for its growth rate (section 2.3.3). In conclusion, we show that size regulation in populations does not require any ability to sense size. Furthermore, the results suggest that size regulation is an intrinsic property of growing and dividing cells.

2.2 Materials and Methods

2.2.1 The Model: Assumptions and Implementations

We constructed a size regulation model with variable G_1, constant S/G_2 duration and an instantaneous mitotic event, i.e. cell division. The model consists of ODEs joined with a stochastic function and is embedded into a multiscale simulation environment (MSE), which is implemented using the programming language *Python* (van Rossum, 1995). The model contains a total of seven ODEs for seven modeled species (Fig. 2.1; Appendix A, Tab. A1) and a total of 8 parameters that may assume different values in G_1 and S/G_2 phase (Tab. 2.1).

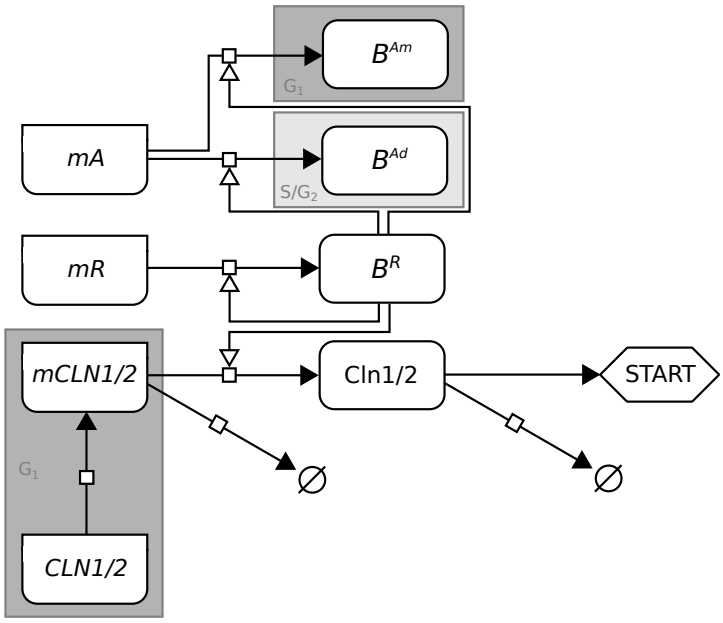


Figure 2.1: **Wiring diagram of the model.** Systems Biology Graphical Notation representation of the model (Le Novère et al., 2009).

We employed the LSODA algorithm (Petzold, 1983) from the FORTRAN library odepack to solve the ODE system. The function *odeint* from SciPy's *integrate* package is a wrapper around the LSODA algorithm. For solving the system the following parameters were used: relative and absolute tolerance $rtol = 10^{-5}$ and $atol = 10^{-6}$, respectively, maximal function evaluations $mxstep = 10^4$, for all others we used default values. In addition to the ODE system, we implemented a stochastic function for $mCLN1/2$ such that the model can account for stochastic transcription (transcriptional bursts) that occurs with a certain probability (40% in all simulations, unless specified otherwise). Other precursor molecules are maintained at a constant number of molecules per cell (Barik et al., 2010). The equations are shown in Table 2.2.

	Specification	G_1	S/G_2
k_{d1}	degradation rate *mCLN1/2*	0.1	0.1
k_{d2}	degradation rate Cln1/2	0.1	10
k_p	production rate Cln1/2	0.35	0.35
k_R	fraction of biomass allocation to internal biomass	4.75	2
k_{Am}	fraction of biomass allocation to structural biomass (mother)	1	0
k_{Ad}	fraction of biomass allocation to structural biomass (daughter)	0	1
k_{growth}	nutritional setup	0.02	0.02
k_x	conversion factor structural biomass to m^2 membrane	1	1

Table 2.1: **List of parameters.** The model has 8 adjustable parameters that can assume different values in G_1 and S/G_2 phase.

Constructing the model we assumed that a cell produces two types of biomass, structural and internal biomass (B^A and B^R). The structural biomass is proportional to the area (A) of a cell (equations 9 and 10) and constitutes cell components that define the outer area of the cell, i.e. the membrane and the cell wall. The total area of a cell is assumed to be the sum of the area of the mother (A^m) and the area of the bud (A^d), if present (equation 11). The total area determines the cellular uptake capacity for nutrients (equation 12). The internal biomass (B^R) is the soluble biomass that is proportional to cellular metabolic capacity. It determines the ability to incorporate intracellular nutrients into new biomass. The total biomass production is dependent on the internal biomass. Therefore, all production reactions (e.g. Cln1/2 translation) are implemented with second order kinetics, dependent on some precursor and the internal biomass (equations 2-5). All entities are computed as amounts (not concentrations) and thus, parameters are scaled with the volume V of the cell to account for volume changes in the calculations, as defined in equation 1.8. As mentioned, this phenomenon does not affect first order reactions. Cells are approximated as spheres, which is why we calculated the volume of a cell according to $V = A^{3/2}$ (equation 13). Consistent with the area calculations, we assumed that the total volume of the cell is the sum of the volume of the mother cell (V^m) and the volume of the daughter cell (V^d) (equation 14). Herein, we neglected the error that is imposed by volume calculations of two intersecting spheres in case of a budding cell.

Furthermore, we assumed that cells may allocate their resources according to cell cycle phase and distribute them to either structural or internal biomass. As the total amount incorporated per time is defined by the metabolic capacity, increase in one leads to decrease of the others. We assume that in G_1 phase of the division cycle only the mother grows, hence $k_{Ad} = 0$, and that the majority of metabolic capacity flows into the internal biomass $k_R = 4.75$ vs. $k_{Am} = 1$, reflecting the high investment in ribosomal RNA and proteins (Warner, 1999; Xiao and Grove, 2009). In accordance with empirical data, we changed these conditions for the S/G_2 phase allocation (Aldea et al., 2007).

1.	$mCLN1/2' = -k_{d1} * f(mCLN1/2, t)$
2.	$\mathrm{Cln1/2}' = f_{flux} * \frac{k_p}{V} * mCLN1/2 * B^R - k_{d2} * \mathrm{Cln1/2}$
3.	$B^{R\prime} = f_{flux} * k_{growth} * (\frac{k_R}{k_R + k_{Am} + k_{Ad}}) * \frac{1}{V} * mB^R * B^R$
4.	$B^{Am\prime} = f_{flux} * k_{growth} * (\frac{k_{Am}}{k_R + k_{Am} + k_{Ad}}) * \frac{1}{V} * mB^A * B^R$
5.	$B^{Ad\prime} = f_{flux} * k_{growth} * (\frac{k_{Ad}}{k_R + k_{Am} + k_{Ad}}) * \frac{1}{V} * mB^A * B^R$
6.	$mB^{R\prime} = 0$
7.	$mB^{A\prime} = 0$
8.	$f(mCLN1/2, t) = \begin{cases} mCLN1/2(t_i) + randint(0,1) & \text{if } t = t_i \\ mCLN1/2(t) & \text{otherwise} \end{cases}$
9.	$\triangle A^m = k_x * \triangle B^{Am}$
10.	$\triangle A^d = k_x * \triangle B^{Ad}$
11.	$A(t) = A^m(t) + A^d(t)$
12.	$f_{flux} \equiv A(t)$
13.	$V^x(t) = A^x(t)^{3/2}$ for $x \in \{m, d\}$
14.	$V(t) = V^m(t) + V^d(t)$

Table 2.2: **List of equations.** The model consists of seven differential equations (1-7), a stochastic function (8) and six algebraic equations (9-14).

Here, we assumed that much less energy goes into the internal biomass ($k_R = 2$) and accordingly, more is allocated to bud growth ($k_{Am} = 0$, $k_{Ad} = 1$). Finally, we reasoned that the nutritional setup that we exposed our cells to (k_{growth}) should have a strong effect on the long term growth, but not directly on Cln1/2 production, to avoid direct nutrient regulation of the cell cycle, which is why we included the growth parameter exclusively in equations 3-5.

2.2.2 A Multiscale Simulation Environment

In order to simulate complex cell cultures, we integrated the model into a multiscale simulation environment, where every cell is implemented as an object that contains predefined attributes. This enabled us to record the total duration of G_1 phase (from birth until Cln1/2 rises above an arbitrary threshold), the cell size (over time, at birth and division), division ratios between mother and daughter and all of the components of the wiring diagram in Figure 2.1 (see also Appendix A, Tab. A1) over time for every individual cell. Thus, a simulation generates a complete pedigree of growing and dividing mother and daughter cells (Soueidan et al., 2007). Division occurs when the cells accumulate 150 arbitrary units of Cln1/2, reflecting a critical localized activity in the nucleus required to trigger START transition. While the nuclear volume has been shown to increase slightly (Jorgensen et al., 2007), the abundant excess of substrates (limitation of Cln1/2) sets the stage for a zero-order ultrasensitivity and hence rapid transition after nuclear accumulation of Cln1/2 (Goldbeter and Koshland, 1981). For simplicity, we implement this as a threshold.

At initialization of the simulation, an adjustable, but fixed number of potential cells (10.000) is setup in a matrix (100×100) and an adjustable number of *in vitro* cells is initialized. Throughout this work, all simulation were initialized with 10 cells, unless explicitly stated otherwise. The initial conditions of those cells (Appendix A, Tab. A1) are chosen such that they resemble newborn cell conditions that emerge during the simulations. The array represents a virtual culture dish with the initial number of cells at time 0. With the resolution of 1 minute time, the cells now grow and divide. To this end, we implemented an algorithm (outlined below), which computes the state of each cell for each time step and stores the information. A general advantage of the MSE and, with it, the positioning of cells in a matrix is that different scenarios of cell to cell interaction can easily be implemented, e.g. nutritional competition, cell adaptation after some type of stress or cell to cell signaling, as e.g. in mating.

2.2.3 Parameter Fitting

For the parameter adjustment of the model we used published single cell experimental data (Aldea et al., 2007). During the fitting procedure values for the parameters k_R, k_{Am} and k_{Ad} have been tuned such that in the simulations the qualitative behavior of the volume increase of a single cell resembles the data (see Fig. 2.4). In addition, k_p and k_{growth} values are chosen such that the mean G_1 duration of daughter cells is approximately 100 minutes.

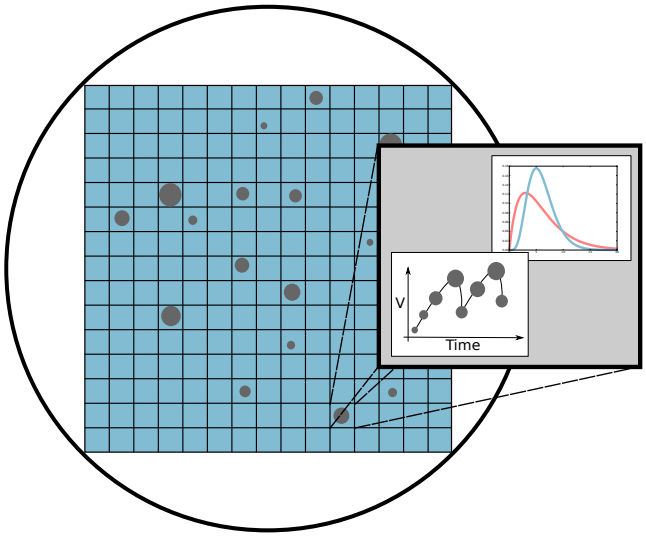


Figure 2.2: **Schematic illustration of the multiscale simulation environment.** Individual cells are followed in parallel and newborn daughters arise as long as the grid contains empty slots. The environment is used to generate a complete pedigree of growing and dividing cells.

Algorithm 1 The multiscale simulation algorithm in pseudo-code.

```
for every time step do
  for every cell do
    if alive and not too old then
      check_age()
      if Cln12 < threshold and not in S/G2 then
        check_transcriptional_burst()
        solve_odes(p =values_G1)
      else if in S/G2 then
        solve_odes(p =values_G2)
      else
        check_space()
        compute_mother_daughter_ratio_according_V^m_V^d()
        distribute_B^R_according_V^m_V^d()
        reset_B^Ad_in_mother()
        reset_A^d_in_mother()
        create_daughter_from_bud(A^d)
      end if
    end if
  end for
end for
```

On a different note, we chose for k_{d2} to display a hundred-fold difference between the two different phases to represent active Cln1/2 protein destruction/inactivation that occurs during S phase of the division cycle (Lanker et al., 1996). Finally, the S/G_2 phase, which is constant, is set to 90 minutes by default. Thus, our *in vitro* conditions resemble *in vivo* conditions that show 190 minutes generation time for a single newborn cell.

2.2.4 Model Validation

Finally, we validated our model using experimental data. Herein, the model has been validated by comparison with independent data that has not been used in the fitting procedure. The results of the model validation are displayed in Figures 2.7 and 2.8. In summary, the model was fitted to single cell data (Aldea et al., 2007) and validated on a population level with population-averaged data from two different sources (Egilmez et al., 1990; Cookson et al., 2009). The model can nicely reproduce single cell behavior and predict population-averaged behavior of a cell culture.

2.3 Results

2.3.1 A Model Linking Growth and Division

We employed a core model to assess the minimal requirement for size regulation in living cells (Fig. 2.3). This core model includes biomass production as a function of present biomass, i.e. a self-replicating system similar to previous work (Molenaar et al., 2009), and surface-to-volume ratio. Division occurs when the activity of a regulatory protein, the G_1 cyclin Cln1/2, reaches a given threshold. There is no regulatory feedback, size sensing, or size regulation. The model is based on two high level assumptions; (i) to grow, cells need to (a) take up nutrients and (b) incorporate nutrients in biomass and (ii) that metabolic efficiency decreases with decreasing area to volume ratio. The model accounts for two qualitatively different forms of biomass: structural biomass (B^A) and internal biomass (B^R). The structural biomass is proportional to the area (A), it includes cell wall and cell membrane and it determines the cellular uptake capacity for nutrients. The internal biomass is set proportional to the cellular metabolic capacity (R) and determines the ability to incorporate intracellular nutrients into new biomass. The cell allocates resources to either structural or internal biomass depending on the cell cycle phase. Furthermore, structural biomass is allocated either to the mother (B^{Am}) or daughter (B^{Ad}) part of the cell and stays with that part, while internal biomass is split at division proportionally to volume.

The model includes a simplified cell cycle with only two phases: G_1, with growth of the unbudded mother and a bias towards allocation to internal biomass, and S/G_2, with polar growth into the bud, relatively small allocation to internal biomass and no growth of the mother. The phases are interspaced by START - at which cells decide to divide - and an instantaneous M phase after a set time delay at which cells divide. All regulation occurs at the level of START, which is triggered by Cln1/2 accumulation in the nucleus.

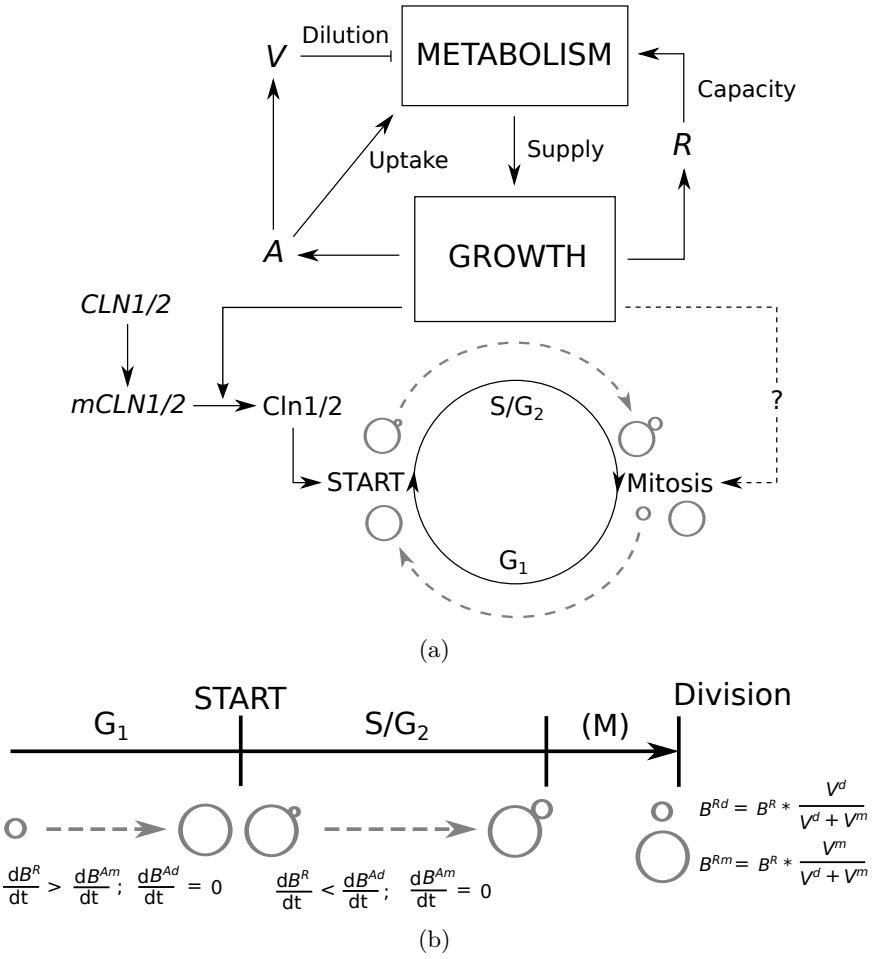


Figure 2.3: **Modeling approach.** (a) Cell cycle is approximated as two growth phases, G_1 and S/G_2, separated by two events: START and Mitosis. Growth allocates resources to two types of biomass: structural biomass (B^A) and internal biomass (B^R). B^A is proportional to surface area (A), which determines volume ($V = A^{3/2}$). B^R is proportional to metabolic (biosynthetic) capacity (R). Metabolism depends on uptake capacity (A) and relative metabolic capacity (R/V) and determines the resources available for growth. Thus, metabolism also determines translation of *mCLN1/2*. Cln1/2 increase in G_1 triggers START transition. Regulation upstream of Cln1/2 was removed and *CLN1/2* transcription is stochastic. (b) Resource allocation differs between two growth phases. In G_1 resources are allocated to mother cell and primarily to B^R. At START, cells polarize and start growing buds by targeted secretion, here described as an altered resource allocation. In S/G_2, resource allocation shifts towards B^A, specifically to the bud (B^{Ad}). S/G_2 duration is constant and Mitosis instantaneously separates mother and daughter cells. Both retain their structural biomass (B^{Am} and B^{Ad}, respectively) and inherit a share of the internal biomass (B^R) proportional to their volume.

1.	nutrient uptake proportional to cell area
2.	transcription is stochastic
3.	biomass production is dependent on the internal biomass (B^R)
4.	thus, all production reactions are implemented with second order kinetics, dependent on some precursor and the internal biomass (equations 2-5)
5.	cells are approximated as spheres, thus $V = A^{3/2}$
6.	cell area is sum of mother and daughter area
7.	cell volume is sum of mother and daughter volume
8.	cells may allocate their resources according to cell cycle stage to either structural or internal biomass
9.	structural biomass can go into area mother (A^m) and area daughter (A^d), separately
10.	in G_1 there is no bud growth
11.	after START only the bud grows
12.	increase of metabolic capacity is strong in G_1 - less after START
13.	threshold for nuclear kinase activity (zero order sensitivity)
14.	there is targeted Cln1/2 destruction/nuclear exclusion after START
15.	cells that are too old go quiet after $\sim$ 24 divisions
16.	S/G_2 is constant
17.	mitotic cell division event is instantaneous
18.	biomass precursors are always available (constant)

Table 2.3: **List of modeling assumptions.**

Importantly, *CLN1/2* transcription is entirely stochastic to reflect the lack of upstream regulatory networks, see sections 1.2.1 and 2.2 for details. In summary, it comprises eight parameters (Tab. 2.1), seven ODEs, a function for stochastic transcription, six algebraic equations (Tab. 2.2) and rests on a set of explicit assumptions (Tab. 2.3).

2.3.2 The Model can Reproduce Characteristic Aspects of the Cell Cycle

The core model was calibrated manually using high resolution literature data on cell growth that distinguishes between mother and bud growth (Aldea et al., 2007). Figures 2.4 and 2.5 show the fit as well as the trajectories of each of the five time-dependent variables in a single cell over two cell division cycles, following only the mother in the second cycle. In this particular case, the newborn daughter spends a long time ($\sim$100min) in her first G_1 phase, before the stochastic *CLN1/2* expression leads to sufficient accumulation of Cln1/2 proteins to trigger START. In the S/G_2 phase, growth is redirected to the bud leading to an accelerated area and volume increase, which eventually leads to a biphasic growth pattern within one cycle. The phase specific alteration of the growth rate is in accordance with experimental observations (Aldea et al., 2007; Cookson et al., 2009; Goranov et al., 2009). Additionally, *CLN1/2* transcription ceases and existing Cln1/2 proteins are actively degraded.

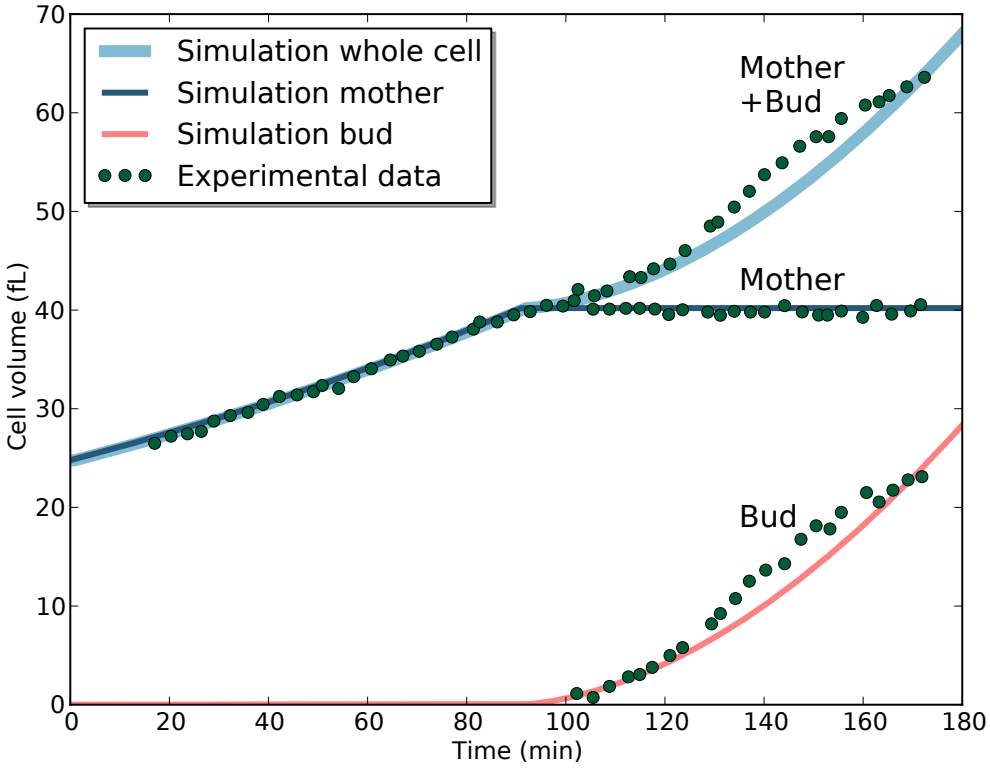


Figure 2.4: **The model captures single cell behavior.** Model parameters were adjusted to *in vivo* growth data quantifying mother and daughter specific growth of budding cells (green dots; Aldea et al. (2007)). The adjusted parameter values (Tab. 2.1) were used for all subsequent analysis. Solid lines indicate the simulation results.

After a set time delay, the cell passes through Mitosis and volume, area and metabolic capacity split between the mother and the daughter (resulting in a drop as only the mother line is followed). The cell enters her second G_1 larger and with higher metabolic capacity, resulting in a faster accumulation of the G_1 cyclins and hence a much shorter G_1 phase, which is in accordance with empirical data (Brewer et al., 1984; Cookson et al., 2009).

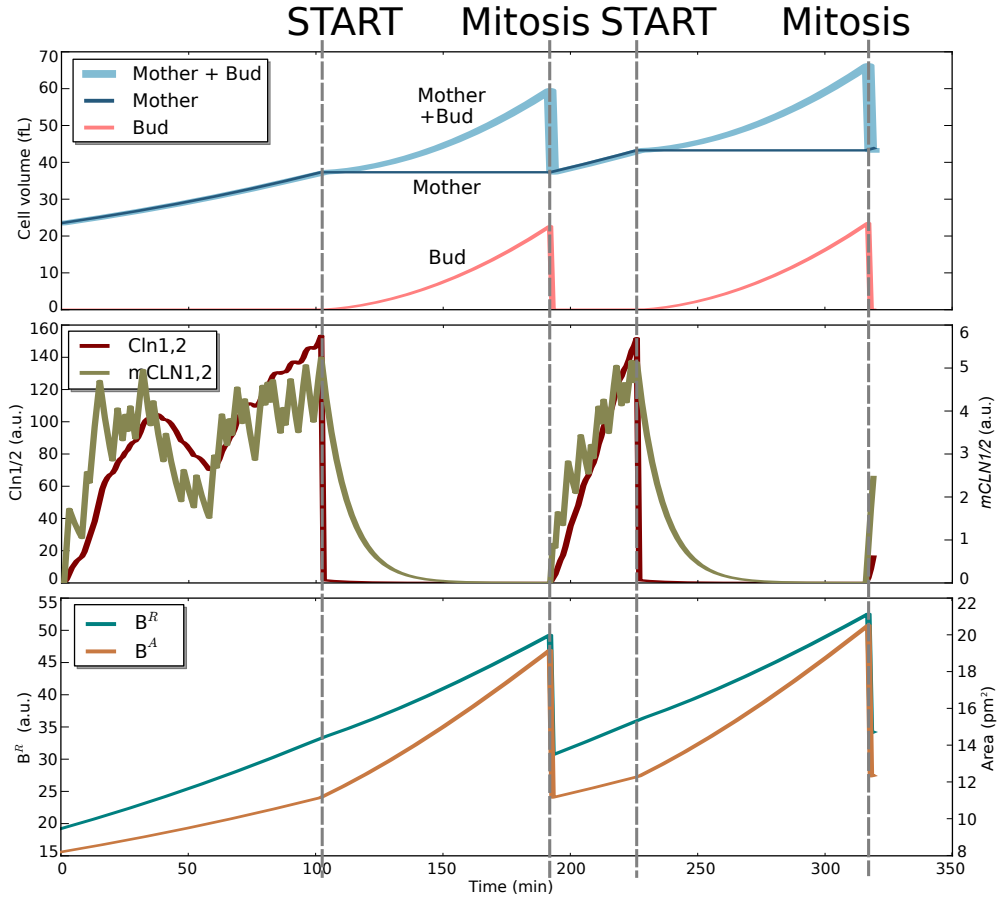


Figure 2.5: **Single cell simulation over two cell cycles.** Displayed are cell volume (upper panel), *mCLN1/2* and Cln1/2 levels (middle panel) and structural and internal biomass (lower panel). Cell volume shows a biphasic growth pattern reflecting the shift in resource allocation from internal to structural biomass. *CLN1/2* transcription is stochastic in G_1 but absent in S/G_2 and Cln1/2 is actively degraded during S/G_2. Note that the second G_1 phase is much shorter than the first.

2.3.3 Size Regulation on the Single Cell Level is not Needed for Population Size Regulation

We proceeded to evaluate the population level predictions of the core model. A culture with ten identical cells has been initiated to generate a complete pedigree stemming from these cells using the multiscale simulation framework (Fig. 2.2; section 2.2). The individual cell lines were followed over several generations on both the mother and daughter lines. Figure 2.6 shows a typical simulation result for such a virtual culture.

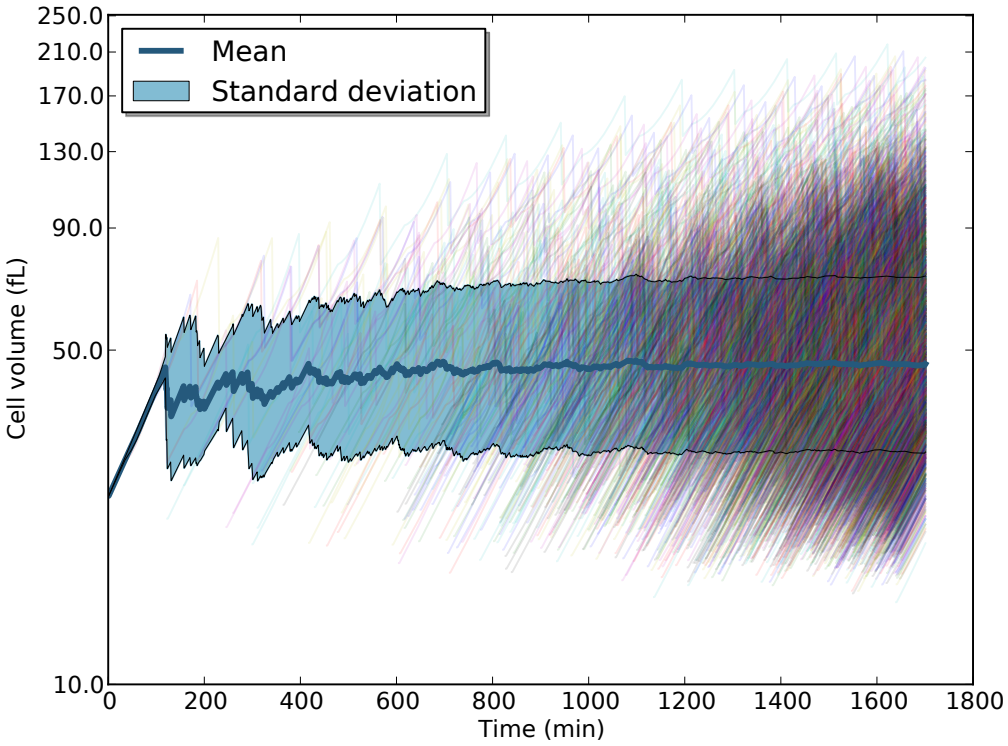


Figure 2.6: **Growth of a complex population.** Starting from ten cells, a complete pedigree of growing and dividing cells with a final population size of 10,000 cells has been generated. Cell volume in fL of individual cells (thin colored lines) and the population average (thick dark blue line) with standard deviation (blue area) are shown. First ten cells start with identical initial conditions (Appendix A, Tab. A1). Subsequent initial conditions are defined by mother and bud at the time of division. Hence, the population quickly falls out of synchrony and size average and variance stabilize.

The thin lines show the individual cell sizes, which increase until division, split in two to follow the mother and daughter individually, and resumes their increase until the next cell division. The bold blue line indicates the population average and the shaded field the span of one standard deviation around the average. The average and

variation in cell size quickly stabilizes despite the fact that the size of individual cells grows strictly monotonously. For an individual cell this leads to an increase in cellular size over generations, which is consistent with *in vivo* observations (Fig. 2.7; Egilmez et al. (1990)).

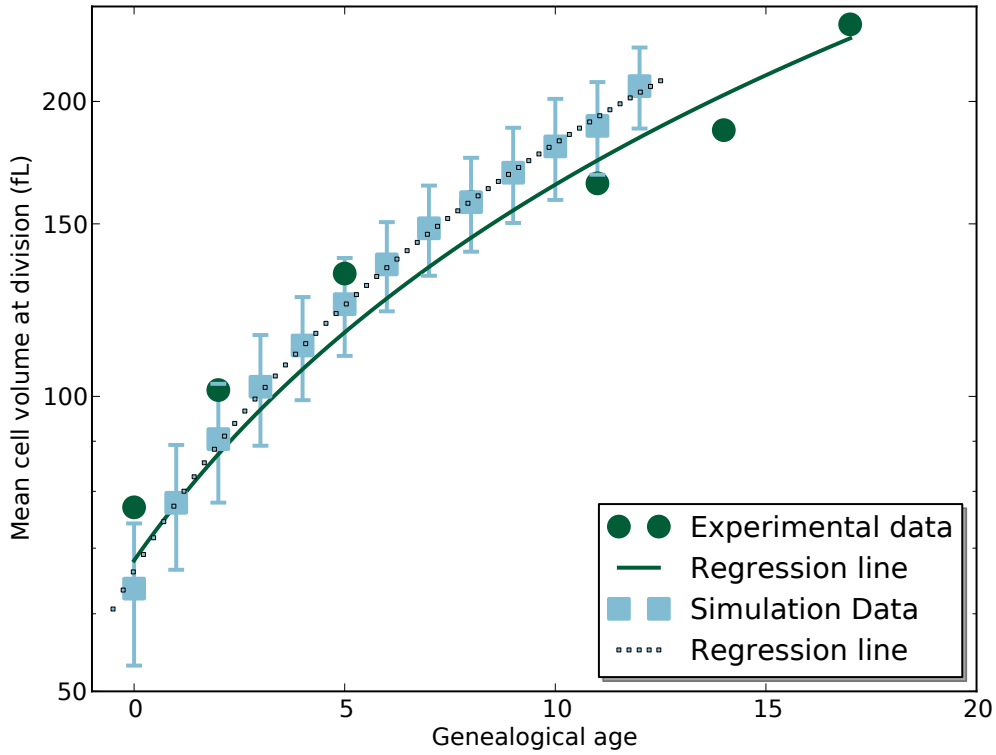


Figure 2.7: **Mean cell volume (fL) at division as a function of generation.** Individual cells grow larger for each generation in simulations (blue) and *in vivo* data (green; Egilmez et al. (1990)), displayed as a good agreement also between the *in vivo* (solid green) and *in silico* (dashed blue) regression lines.

The simulation predicts that the older (and larger) mothers will progress faster through G_1. This is consistent with empirical data, although the *in vivo* decrease in G_1 duration stagnates after only a few generations (Fig. 2.8 (a); Cookson et al. (2009)). Finally, the model accurately predicts that older mothers will retain a larger fraction of the total volume, although again the *in vivo* effect saturates after the first few generations (Fig. 2.8 (b); Cookson et al. (2009)).

Note that none of the *in vivo* data presented in Figures 2.7 and 2.8 was used for model fitting. Taken together, the core model is in excellent agreement with empirical observations and suffices to provide size regulation on the population level without any size regulation on the individual cell level.

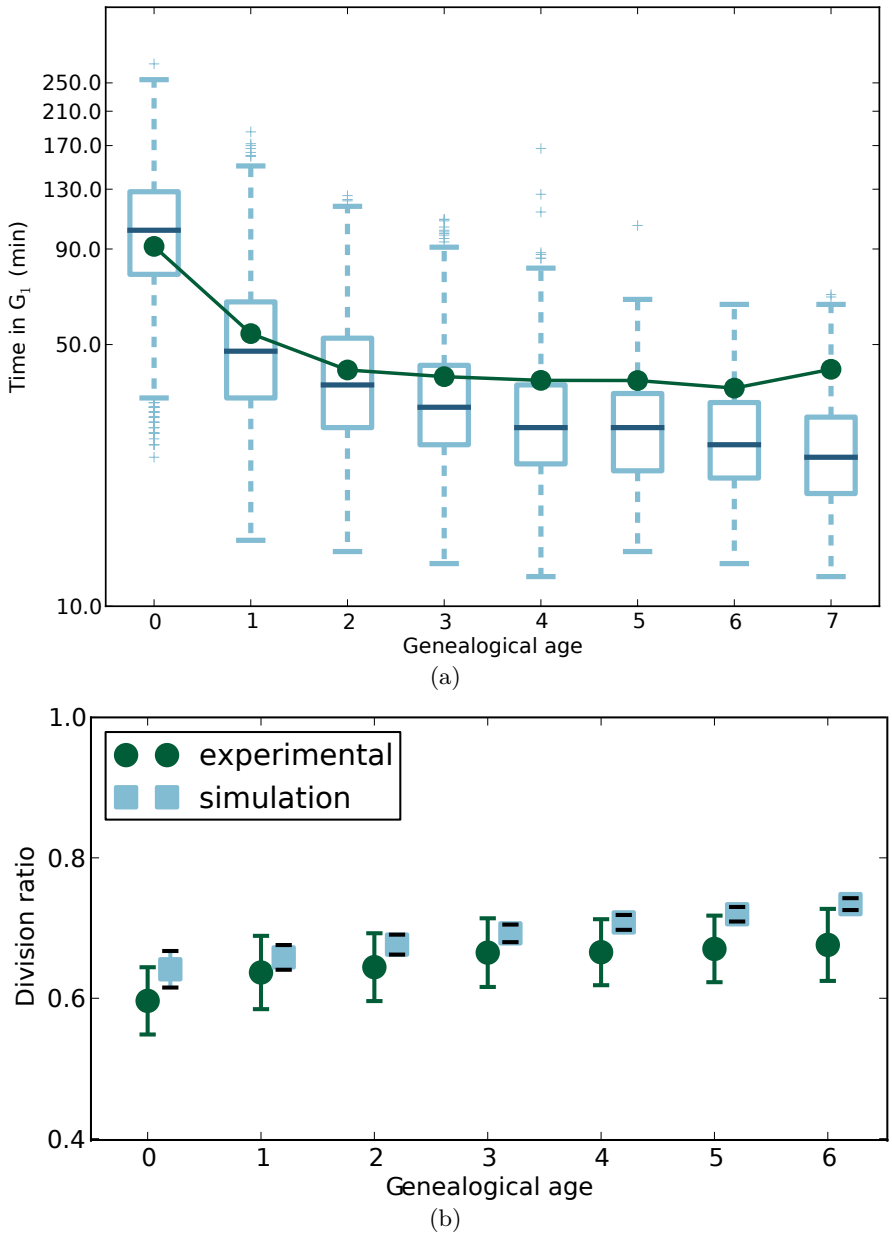


Figure 2.8: **G_1 duration decrease (a) and fraction of volume retained by mothers at division (b) with increasing age.** (a) The box plot displays G_1 duration in simulations (blue) compared to experimental data (green; Cookson et al. (2009)). In simulations and experimental data, older and, hence, larger cells pass through G_1 faster. (b) Fraction of volume retained by mothers at division as a function of age in the model (blue) and *in vivo* (green) (Cookson et al., 2009).

2.3.4 The Model Captures Growth Rate Specific Population Behavior and Suggests that Effective Size Regulation over Different Growth Rates Requires a Variable (Rate-Adapted) G_2 Duration.

A realistic growth model should also be able to capture the change in growth rate and cell size distribution associated with different growth media. It is well known that an increased growth rate gives larger cells *in vivo* (Johnston et al., 1979; Tyson et al., 1979). While the core model is too abstract to account for the exact media composition, it includes a "nutrient quality" term that allows for different growth rates (see section 2.2 for details on k_{growth}). We simulated four different nutritional conditions with mass doubling times ranging from $\sim$ 2 hours ($k_{growth} = 0.04$) to $\sim$ 4 hours ($k_{growth} = 0.01$) and compared the size distribution (Fig. 2.9).

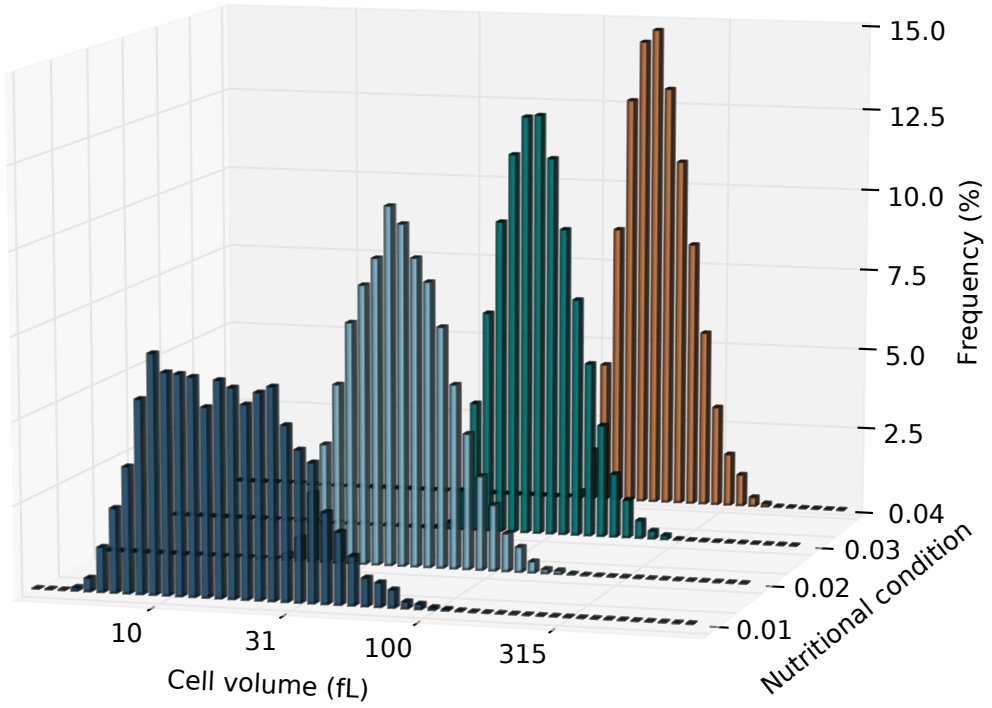


Figure 2.9: **Population distributions of cells growing at four different growth rates**, from low (dark blue) to high (orange). Note that cell sizes become larger and less variable as the growth rate increases. The cultures had mass doubling times of 3.96 hours (0.01), 2.68 hours (0.02), 2.19 hours (0.03) and 1.96 hours (0.04).

Note that the model predicts both a larger size and decreased biological noise in the faster growing cultures. The faster growth rate also leads to a shorter G_1 duration (Fig. 2.10). However, at very fast growth rates, the core model loses the asymmetry between mothers and daughters.

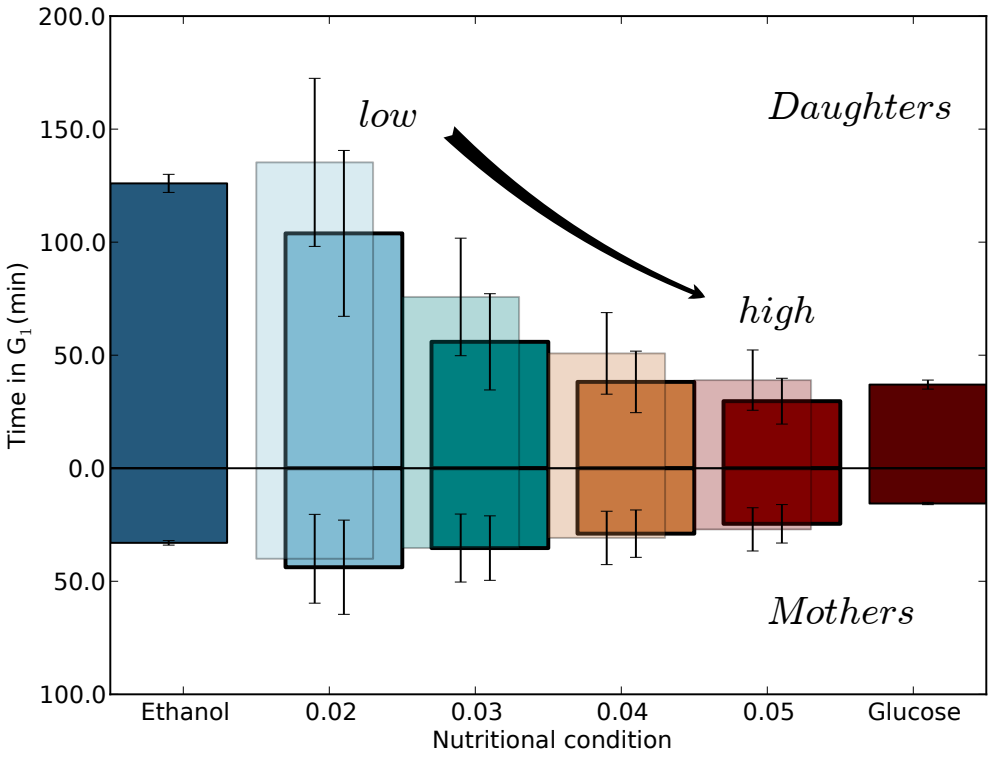


Figure 2.10: **Time in G_1 as function of growth rate** for daughters (upper bars) and mothers (lower bars), and compared to experimental data on cells grown on glucose (dark red) or ethanol (dark blue; Di Talia et al. (2009)). The front bars outlined in black display G_1 duration with a S/G_2 time of 90 minutes (our standard conditions) and the faded bars behind display G_1 duration with a S/G_2 duration of 60 minutes.

It seems that the assumption of a constant S/G_2 duration over different growth rates is too simplistic, because simulations show that the mother/daughter asymmetry can be partly restored by altering the S/G_2 duration (Fig. 2.11). Indeed, empirical data supports the observation that G_2 duration changes over different growth rates (Barford and Hall, 1976). Taken together, the core model qualitatively captures the effect of altered nutrient conditions very well but it appears that regulation of the S/G_2 phase duration is required to fully describe adaptation to altered growth conditions.

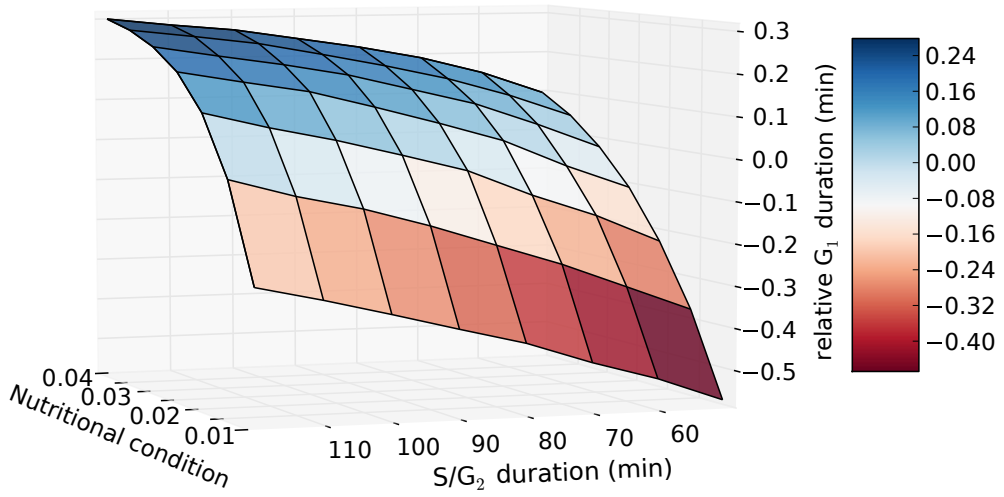


Figure 2.11: **Effect of S/G_2 duration on asymmetry in G_1 duration** between mothers and daughters at different growth rates. The surface plot represents the $\log_{10}$ ratio between the G_1 durations for mother/daughters as function of growth rate and S/G_2 duration. Note that changes in either growth rate or S/G_2 duration alone leads to change in the mother/daughter ratio and that both must be adjusted to retain the same relationship.

2.3.5 Average Cell Size Converges to a Point Attractor, that is Characteristic for a Given Growth Rate

The above results are consistent with average cell size converging to a stable point attractor. Figure 2.12 shows that different growth rates result in defined population growth patterns similar to the reference simulation (Fig. 2.6). An initially synchronous population rapidly desynchronizes and the population average and variance converge to a growth rate specific level. Furthermore, shifting between qualitatively different media resets the size that is specific for the growth rate (Fig. 2.13), as experimentally observed (Johnston et al., 1977). To explore the nature of this attractor, we proceeded to test the impact of initial conditions. Unlike growth media composition, initial conditions should not affect the final size distribution.

As shown in Figure 2.14, the point attractor is stable against perturbation in initial levels of structural and internal biomass by at least two orders of magnitude. In all cases, the proliferating population rapidly converges to the size average determined by the growth rate.

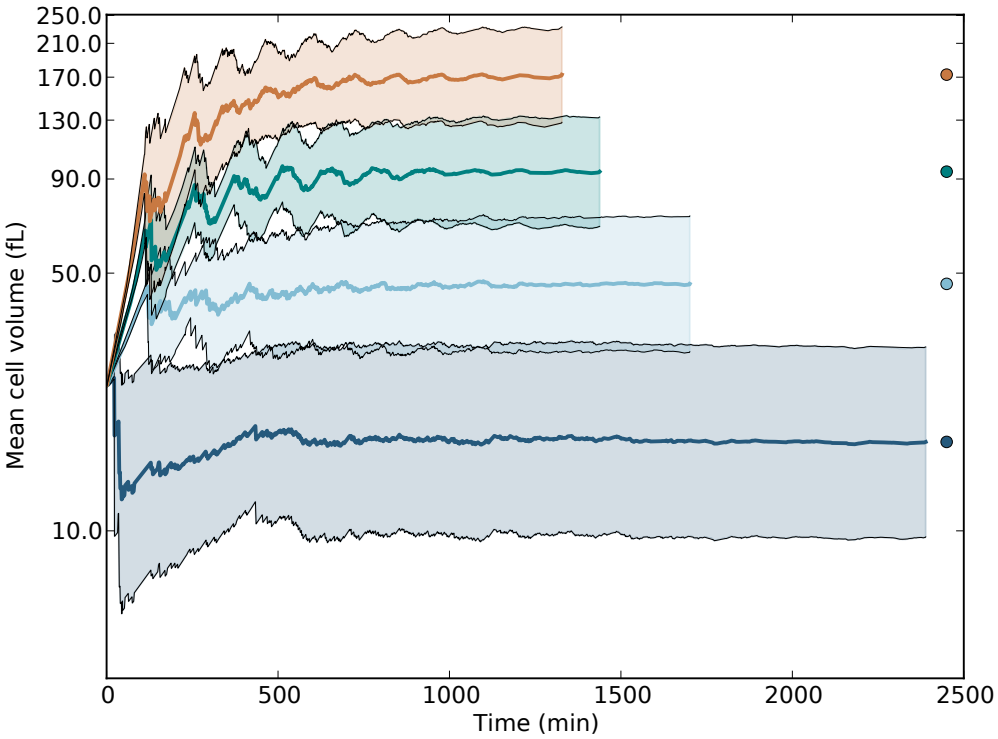


Figure 2.12: **Population size average** (solid lines) and the span within one standard deviation (shaded areas) for four different growth rates (see Fig. 2.9) over time. Despite identical initial condition, average size stabilizes on different levels depending on specific growth rate. Size average increases and variance decreases with increasing growth rates.

Finally, we examined to what extent this size regulation could be tied to the set S/G_2 time. To simulate a noisy S/G_2 duration, we sampled from a normal distribution with increasing standard deviations around a mean of 90 minutes ranging from 0 to $\sim$ 10 % of the mean on a log scale, resulting in S/G_2 durations in the range of 20 up to easily 250 or more minutes (Appendix A, Fig. A1). The resulting simulation shows that while the variability increases, the attractor remains stable but the point of attraction changes somewhat in every simulation. Intriguingly, we see that increasing variability in the S/G_2 duration leads to decreased average cell size (Fig. 2.15). This may be related to an altered population distribution, as the simulated culture's population structure deviates from the theoretical population structure with respect to percentage of cells in different genealogical ages (Appendix A, Fig. A2). Overall, the model validation shows

that the average size converges to a stable point attractor that is robust against altered growth conditions, perturbations in initial conditions and noise in S/G_2 duration.

Summarizing our results, we compiled a comprehensive list of characteristic yeast cell cycle/growth aspects with references to the data source, shown in Table 2.4. It is noteworthy, that the minimal model can account for nearly all points of the reference list, indicating that size regulation is a systemic property of growing and dividing cell populations, instead of the result of a size sensing mechanism.

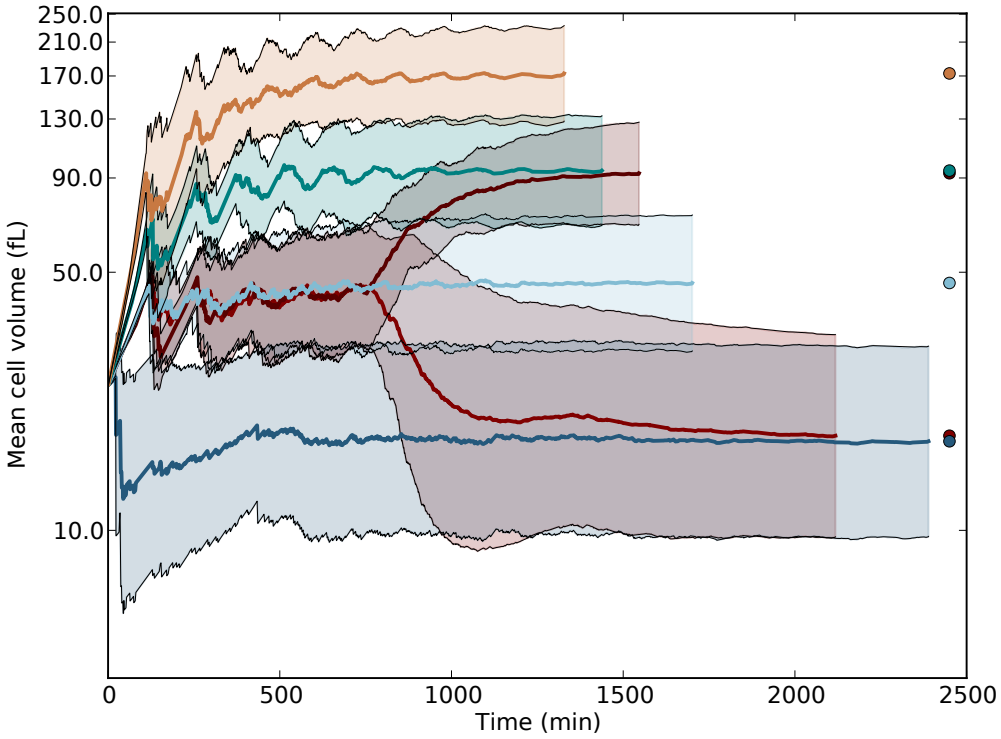


Figure 2.13: **A nutritional shift** from growth rate 0.02 to 0.01 or 0.03 (dark red lines) resets the specific average cell size and standard deviation characteristic for the new growth rate. Continuously growing cultures are identical to those in Figure 2.12.

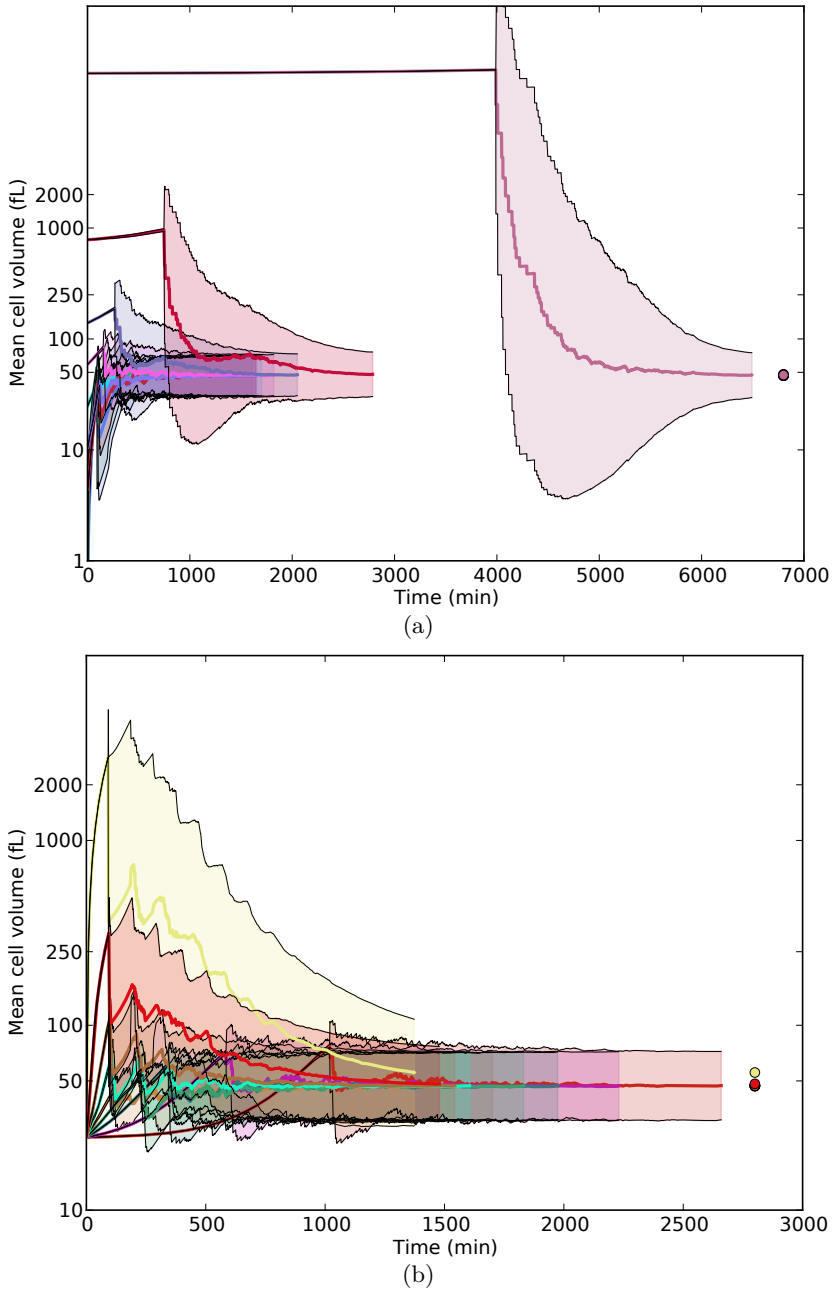


Figure 2.14: **Effect of structural (a) and internal (b) biomass perturbations** on size distribution by two orders of magnitude in each direction. (a) Perturbation of B^A results in START transition delay with increasing initial size. (b) The growth rate specific attractor is stable against perturbation of B^R. In (a) and (b) average population size rapidly converges to a growth rate characteristic size.

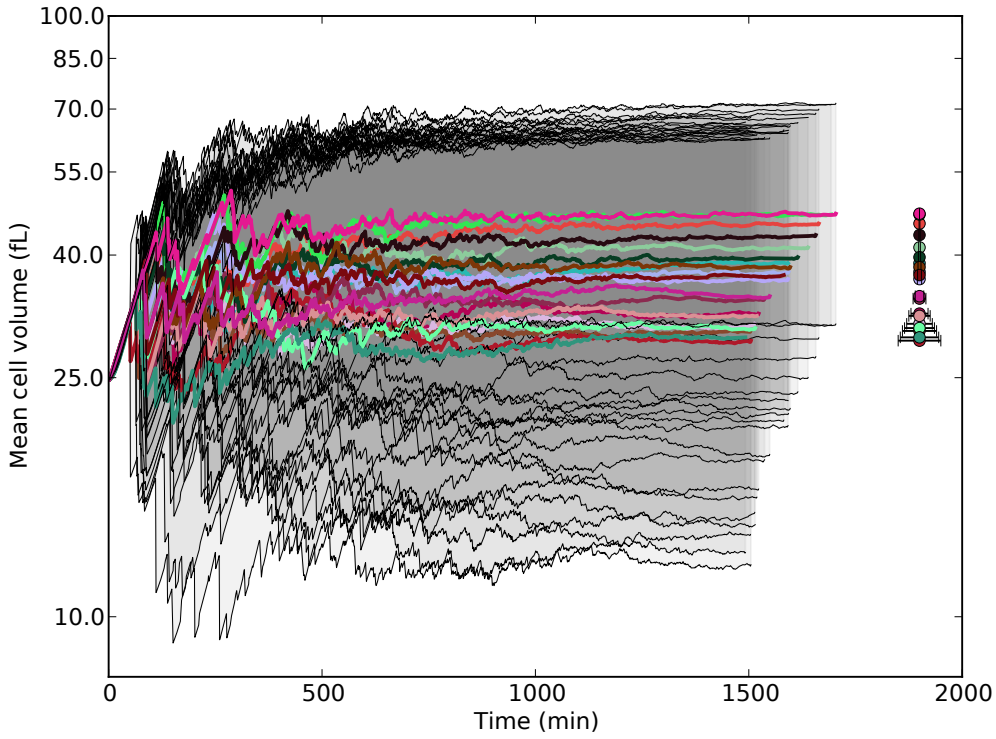


Figure 2.15: **Effect of noise in S/G$_2$ duration.** The set 90 minutes S/G$_2$ duration was replaced with a random duration based on a log normal distribution around 90 minutes with increasing standard deviation. The attractor remains stable and the average cell size decreases with increasing S/G$_2$ noise.

2.4 Discussion

As a contribution to the long standing discussion about how cells can sense their size necessary for cell division, we show that a minimal core model without mechanistic size regulation suffices to reproduce the cell growth and division pattern on the single cell level as well as on the population level over a range of growth conditions. The core model abstracts the cell division cycle to two phases separated by two events, cellular composition to two qualitatively distinct types of biomass, growth to uptake and metabolism, and it links growth to cell division by stochastic transcription and translation of a single regulatory protein. We chose this level of abstraction to accurately describe the basic relationship of growth and division, while keeping the computational cost at a minimum at the same time. Most of the abstractions that we employ, except for constant S/G_2 duration over different growth rates, survives the model validation. The metabolic model assumes (i) that the nutrient supply is defined by the uptake, which is proportional to cell surface area, (ii) that its incorporation into biomass relies on the metabolic capacity and (iii) that the efficiency of the incorporation decreases with volume. The metabolic part of the model is a self-replicating system and as such, it is compatible to the one presented by Molenaar et al. (2009), although in essence its structure is even simpler since in our work metabolic capacity comprises ribosomes as well as metabolic enzymes. While there is convincing evidence that the allocation between metabolic enzymes and ribosomes alters with growth rate, we found the distinction between the two superfluous for this model (Goelzer and Fromion, 2011). The difference in allocation over the cell division cycle builds on experimental evidence and the allocation parameters have been adjusted to fit experimental data (Aldea et al., 2007). While the zero allocation to the mother in S/G_2 is likely to be an approximation, there is no significant size difference between mother cells with large and small buds, strongly arguing that mother growth during S/G_2 is insignificant (Hartwell and Unger, 1977). Similarly, the allocation to R (Tab. 2.1) is probably an underestimation at higher growth rates, as up to 80% of the transcriptional machinery in *S. cerevisiae* is dedicated to synthesis of ribosomal components (Warner, 1999; Xiao and Grove, 2009). Despite its simplicity, the growth model realistically describes growth on the single cell level, both over time within a cell division cycle (Figs. 2.4 and 2.5) and over generations (Fig. 2.7).

The cell cycle implementation in our model reflects the objective to analyze size regulation in the G_1 phase. Hence, it includes the isotropic and apical growth phases but excludes the M phase. Furthermore, it excludes DNA replication and hence considers a joint S/G_2 phase. The START transition from G_1 to S/G_2 is implemented as a threshold level of Cln1/2 and hence Cdk1-Cln kinase activity. This is set to reflect localized activity on several distinct targets which require multiple phosphorylations (in accordance with Barik et al. (2010)). The limitation of Cln1/2 and hence active kinase and the excess of substrate sets the stage for zero order ultrasensitivity that is abstracted as a threshold level (Schneider et al., 2004). That the threshold is given in amount reflects that the active kinase is targeted to specific subcompartments that expand slower than the total cell volume (Jorgensen et al., 2007). Once the threshold is reached, the phase transition is considered irreversible despite the lack of positive feedback due to the inhibition of Sic1

and release of Cdk1-Clb5/6 and initiation of DNA replication. The implementation of stochastic transcription leads to a faster loss of synchrony but does not alter the behavior of the cell cultures (Appendix A, Fig. A3). While definitely including the regulatory network underpinning the cell division cycle, the transition mechanism contains none of the components implicated in size regulation and cannot be triggered by increased size (Fig. 2.14 (b)). Finally, the S/G_2 phase is considered of constant duration in accordance with the hypothesis that size regulation occurs only in G_1 in *S. cerevisiae*. As shown above, this assumption does not entirely hold, as also G_2 duration has to be altered to maintain the mother/daughter asymmetry at high growth rates. Despite this limitation, it accurately predicted key properties on the population level, including convergence to a stable average size despite constant growth of single cells (Fig. 2.6-2.8) and the effects of increased size of cells that grow on more favorable nutrients sources (Figs. 2.9, 2.12 and 2.13).

While a model on this level of abstraction is clearly insufficient for detailed molecular conclusions, it allows us to re-evaluate a number of conclusions from previous modeling efforts. First, as mentioned above, the assumption that G_2 duration is constant over different growth rates is an approximation that needs reconsideration. In our model, the mother/daughter asymmetry requires adaptation of S/G_2 duration at high growth rates (Fig. 2.11), which could have different reasons. One of them is probably the fact that our cells lack the daughter specific transcription factors Ace2 and Ash1, which suppress Cln3 transcription and thus, provoke a daughter specific delay in G_1 even at high growth rates (Di Talia et al., 2009). However, experimental data also shows that G_2 length indeed varies in different media, giving rise to the hypothesis that both G_1 and G_2 exit have active "size" regulation (Barford and Hall, 1976). Hence, also describing G_2 regulation will require integration of growth and proliferation in the future. Second, it can help explain the apparent paradox that *CLN3* overexpressing cells are smaller but divide faster (Hall et al., 1998; Barberis et al., 2007). While this remains an apparent paradox when the two growth parameters (size and doubling time) are independent, it may be resolved when considering that the growth parameters are intrinsically antagonistic. In other words, given a set nutrient availability, a decreased time in G_1 will always lead to a decreased size accumulation during the cell division cycle. This would hold in the mother cells even if the G_2 phase was prolonged to compensate for the loss in the first generation, as the growth of mothers is negligible during the S/G_2 phase even *in vivo* (Hartwell and Unger, 1977). Here, we see that these theoretical predictions and empirical observations can be reproduced with our core model (Appendix A. Fig. A4), reinforcing the conception that the metabolic power rather than size triggers START transition and that cells grow larger on richer media because the increase in growth supersedes the decrease in G_1 duration (Hall et al., 1998). Third, we can explore the mechanism underpinning the apparent size regulation. It is important to note that most available models are unsuitable for this purpose, as they explicitly or implicitly set the size distribution, e.g. by linking START to a critical size or by assigning a set division ratio. In the model presented here, both the size and division ratios are outputs and thus, emergent properties of the dynamic system. Based on these results, we agree with the notion that size regulation is an emergent property (Barberis et al., 2007).

	Yeast fact	Reference	
1.	asymetric division	Hartwell and Unger (1977)	√
2.	grow at different rates in different cell cycle phases	Aldea et al. (2007); Goranov et al. (2009); Cookson et al. (2009)	√
3.	grow faster (higher growth rates) at better nutritional conditions	Tyson et al. (1979)	√
4.	cells grow bigger at high growth rates	Tyson et al. (1979)	√
5.	cells grow smaller at low growth rates	Tyson et al. (1979)	√
6.	shifting between different nutrient conditions resets size threshold specific for growth rate	Johnston et al. (1977)	√
7.	most G_1 network mutants are viable and still exert size regulation	Enserink and Kolodner (2010)	√
8.	difference in mother and daughter cell cycle length is established mainly in G_1	Brewer et al. (1984)	√
9.	single cell gets larger with age	Egilmez et al. (1990)	√
10.	older (and larger) mothers will progress faster through G_1	Brewer et al. (1984); Cookson et al. (2009)	√
11.	older mothers will retain a larger fraction of the total volume	Cookson et al. (2009)	√
12.	S phase is constant over wide range of growth rates	Barford and Hall (1976); Johnston et al. (1980); Brewer et al. (1984)	√
13.	G_2 is constant for specific growth rate, but varies with different rates	Barford and Hall (1976)	-
14.	single cell growth is linear/bilinear/exponential	Aldea et al. (2007); Di Talia et al. (2007); Cookson et al. (2009); Goranov et al. (2009)	√
15.	growth after START goes mainly into bud	Aldea et al. (2007)	√

Table 2.4: **List of yeast cell cycle and growth characteristics.** Qualitative statements about growth and cell cycle related characteristics for budding yeast have been assembled. Indicated are those that our model can account for. Although, we do not claim that the data is complete, it provides an overview over the current understanding of properties regarding this system. The list could potentially serve as a comprised reference for future modeling studies.

However, our results reject the hypothesis that it emerges from the regulatory network, which has also been shown to be dispensable *in vivo* (Enserink and Kolodner, 2010), and favor the argument that it emerges as a result of metabolic capacity of a cell (Jorgensen et al., 2004; Rudra and Warner, 2004). In essence, the apparent size homeostasis could even be a side effect of the insurance that the cells will be able to complete the cell division cycle, which would explain the wide tolerance for cell size. Taken together, the coarse grained core model captures cell growth and the cell division cycle surprisingly well and accurately predicts both basic cell behavior on the single cell and on the culture level, as well as emergent properties observed *in vivo* and not yet fully explained theoretically (Tab. 2.4).

The core model presented has low computational costs allowing simulation of relatively large cell cultures, which will facilitate further analysis of the cell division cycle and the regulatory networks surrounding it. The MSE that we present here runs independently of the type of single cell model that is used for the ensemble modeling. In this study, we use a minimal ODE model, which could easily be extended in the future to contain other, more detailed cell cycle regulatory circuits or metabolic components. Also, other already published models could potentially be plugged into the MSE to broaden their scope from single cell to population behavior. We are convinced that the core model and the MSE simulation platform will prove a valuable tool for the cell division cycle modeling community. A problematic point in population simulations is the fast increase of the computational costs, which naturally increases at the same rate as the number of individuals in the population, exponentially in this case. Therefore, we tried two different approaches to enhance the computational performance. First, we implemented a version that caches the results of the ODE solver function evaluations in a temporary memory. In case the function is called again with the exact same input parameters, the results are instantly available and do not have to be computed twice. Unfortunately, the function input in our case is too diverse to save time (gain of speed vs. loss of accuracy). Nonetheless, we kept the architecture in the source code, so that it is free to be used in other applications e.g. in a deterministic version of our model or for different models in which the input to ODE solver function is always similar or rounded. In such a case the cached version could be an advantage that would allow for the simulation of very large cultures. Second, we envisioned multicore usage for computing the population. However, in this case as well, we could not gain an advantage in our simulations, since our single function evaluations are generally too quick. It is only worth it when a parallel function evaluation is faster than time lost in the distributing process of the task to the different cores.

Our most intriguing conclusion is that size regulation as observed *in vivo* can be explained without any ability to sense or regulate size on the single cell level. The concept that metabolic power gates the cell division cycle is not entirely new (Jorgensen et al., 2002; Cook and Tyers, 2007) and would help explain observations such as partial cell division cycle synchronization in cell cultures with strong metabolic oscillations (Murray et al., 2007). It is also consistent with the theory around the rejected size regulatory network, as gating through metabolic power would ensure that cells posses the resources required to successfully complete S phase before they pass START. Hence, it is largely

consistent with previous work, but it also leaves the question of why the sophisticated regulatory network upstream of START has evolved. This could simply be a noise reduction mechanism as has been proposed earlier (Rupes, 2002) or an additional modulator for fine-tuning daughter/mother specific regulation (e.g. differential expression of genes *ACE2* and *ASH1* (Di Talia et al., 2009)). Our simulation results lend partial support to this hypothesis, as the loss of mother/daughter asymmetry at high growth rates might partially be regained when we include differential regulation of mothers and daughters. Furthermore, it also suggests that noise stabilizes together with the mean after relatively few generations and that the diverging extremes are too few to have significant impact on the population. Another appealing hypothesis is that the network evolved to allow the cell cycle to be regulated by additional factors beyond nutrients, most importantly pheromones and environmental perturbations such as dehydration (Escoté et al., 2004; Peter and Herskowitz, 1994). The G_1 arrest is critical to synchronize mating cells and hence ensure the genetic integrity of the resulting zygote (Peter and Herskowitz, 1994). Likewise, loss of turgor is devastating for the cells ability to grow and expand (Chowdhury et al., 1992). Both these signals have dedicated MAP kinase pathways acting on dedicated Cdk1 inhibitors. The pheromone response pathway arrests the cell cycle in G_1 *via* Far1, while the High Osmolarity Glycerol pathway arrest the cell cycle in G_1 *via* Sic1 (Escoté et al., 2004). Hence, the main purpose of the G_1 regulatory network may not be to advance the cell division cycle, but to allow for stable cell cycle arrest in G_1 and to ensure that the arrest can be lifted when conditions are suitable.

In conclusion, we have developed a core model to determine the minimal regulatory network required for size regulation in *S. cerevisiae*. Surprisingly, we find that a model without any such network can explain the *in vivo* size regulation, clearly rejecting the hypothesis that size regulation on the population level requires size sensing and/or regulation on the single cell level. In addition, our results support the notion that growth regulates the cell division cycle in both G_1 and G_2 also in *S. cerevisiae*. Taken together, our results provide a framework to further study the function of the G_1 regulatory network and other cell division cycle questions in a population-oriented manner.

3 A Model for the Spatiotemporal Organization of DNA Replication

In the following chapter, I present a computer model for the DNA duplication procedure in budding yeast. The model is used to study the spatiotemporal organization of the replication process with main focus on the impact of differential origin firing patterns. The chapter is based on:

T. W. Spiesser, E. Klipp and M. Barberis. A model for the spatiotemporal organization of DNA replication in *Saccharomyces cerevisiae*. *Molecular Genetics and Genomics*, 282(1):25-35, 2009.

3.1 Introduction

DNA replication in eukaryotes is considered to proceed according to a precise program in which each chromosomal region is duplicated in a defined temporal order. However, recent studies reveal an intrinsic temporal disorder in the replication of yeast chromosome VI. Here, we provide a model of the chromosomal duplication to study the temporal sequence of origin activation in budding yeast. The model comprises four parameters that influence the DNA replication system: (1) the lengths of the chromosomes, (2) the explicit chromosomal positions for all replication origins as well as (3) their distinct initiation times and (4) the replication fork migration rate. The model and parameter details are outlined in section 3.2. The designed model is able to reproduce the available experimental data in form of replication profiles, as shown for the wild type in section 3.3.1 and for a *clb5*$\triangle$ mutant in section 3.3.2. The dynamics of DNA replication was monitored continuously during simulations of wild type and randomly perturbed replication conditions. Severe loss of origin function showed only little influence on the replication dynamics (3.3.3), so systematic deletions of origins (or loss of efficiency) were simulated to provide predictions that could be tested experimentally. The results of the simulations are shown in section 3.3.4 and discussed in section 3.4. In conclusion, the simulations provide new insights into the complex system of DNA replication, showing that the system is robust to perturbation and giving hints about the influence of disordered firing.

3.2 Materials and Methods

3.2.1 Model Characteristics and Available Data

1. *DNA units.* In the model, a DNA unit (u) is defined as a 500 bp block of DNA. Hence, in the simulation each chromosome is composed of a series of DNA units, corresponding to its original size (L_{org}) divided by 500 to yield the internal resolution size L_{res}. To acknowledge the correct size of the chromosomes, L_{res} is always rounded up. The size of the DNA units (500 bp) defines the resolution of the simulation. The size of the chromosomes was obtained from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000; Kanehisa et al., 2006, 2008).

2. *Origin location.* The location of the replication origins on the chromosomes is sequentially predetermined (Newlon and Theis, 1993). The 11 bp region (ACS) can be found within every 200 bp sequence that exhibits origin activity in the budding yeast (Theis and Newlon, 1997). The chromosomal locations of the replication origins can be found in the *S. cerevisiae* OriDB database, version 1.1.1 (Nieduszynski et al., 2007).

3. *Origin initiation.* Initiation times have been assessed for origins of replication (Raghuraman et al., 2001; Yabuki and Terashima, 2002). They are assembled in the *S. cerevisiae* OriDB, version 1.1.1 database (Nieduszynski et al., 2007). In this work we consider the initiation times provided by a heavy:light (HL) timing study (Raghuraman et al., 2001). The initiation time distribution is shown in Appendix B, Figure B5.

4. *Fork migration rate.* The replication bubble grows bidirectionally and both replication forks migrate at a certain rate (v). According to the data reported in Raghuraman et al. (2001), fork rates range from 0.5 to 11 kilo bases (kb)/minute, with a mean of 2.9 kb/minute and a median of 2.3 kb/minute (Fig. 1.3). Similar mean values were obtained in different studies: 2.8 $\pm$ 0.1 kb/minute (Yabuki and Terashima, 2002) and 3.7 kb/minute (Rivin and Fangman, 1980). In this model we assume that the forks migrate constantly throughout S phase at an approximate rate of 3 kb/minute.

The *S. cerevisiae* OriDB, version 1.1.1 database (Nieduszynski et al., 2007) contained 732 replication origins target sites at the time (July 10^{th}, 2008), approximatively 60% (454) of which are considered in this work. The selection is based on the availability of both chromosomal location and firing time (derived from the HL analysis) for every replication origin. A complete list of the replication origins, the location on the chromosomes and the firing times used in this work is published as a supplementary Table in Spiesser et al. (2009).

3.2.2 The Spatiotemporal Model

Figure 3.1 illustrates the model and its parametrization. As described above, the DNA is divided into units of equal length (500 bp). A two-dimensional array element (Ar) of size L_{res} is assigned to every chromosome. Additionally, two DNA units are added to Ar, introducing artificial boundaries, accounting for the left (Ar_0) and right ($Ar_{L_{res}+1}$) end of the chromosomes.

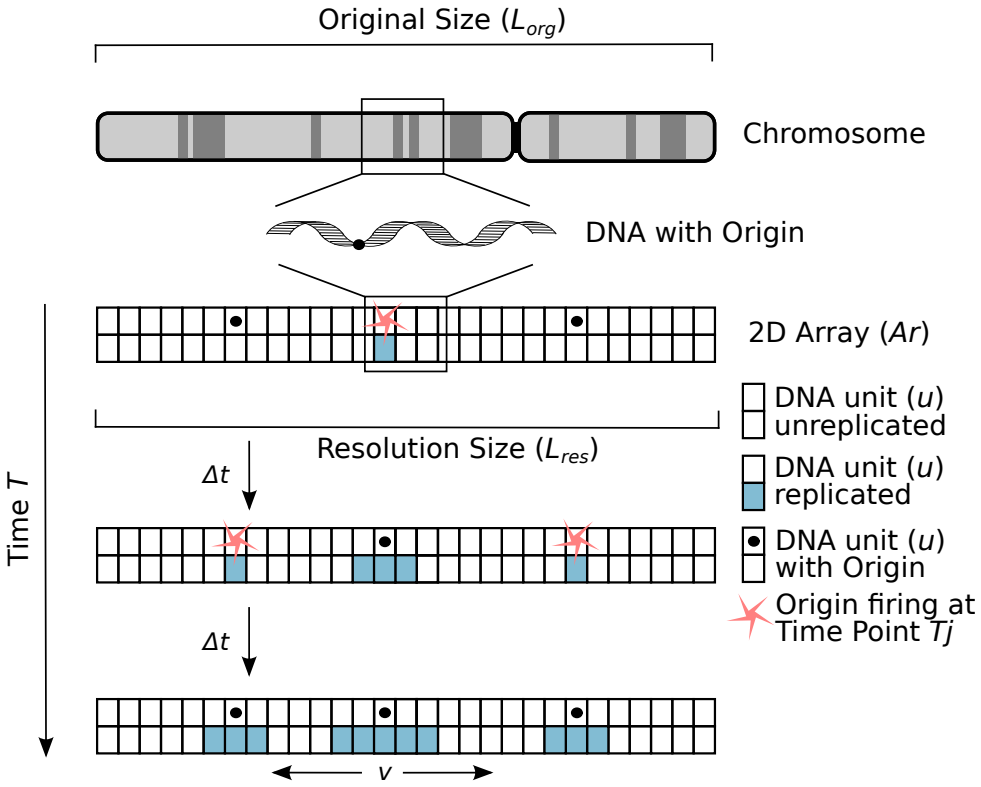


Figure 3.1: **Scheme of the chromosomal duplication model and its parametrization.** The features and the algorithm are explained in the main text.

The array element Ar contains all discrete DNA unit positions ($Ar_{(0:L_{res}+1)}$) and the status of the replication for the position. This is represented by a Boolean Variable, which is set "FALSE" by default indicating that the DNA has not been replicated at this position yet and set "TRUE" only at the end positions of the chromosomes. Another two-dimensional array element (O) stores origin information: origin name, origin position on the virtual chromosome Ar, origin activation time in seconds and the origin activation status, a Boolean Variable, set "FALSE" by default, indicating that the origin has not been activated yet. A variable T represents the replication time. T is the sum of all

discrete time steps t_i, with $(i \in (1, n))$

$$T = \sum_{i=1}^{n} t_i, \tag{3.1}$$

where n is the number of discrete time steps needed to complete DNA replication. One time step equals the time (Δt), that the replication fork needs to go through one DNA unit (Δu). Hence:

$$\Delta t = \frac{\Delta u}{\Delta v}, \tag{3.2}$$

with $\Delta u = 500$ bp and $\Delta v = 3,000$ bp/minute and therefore

$$\Delta t = \frac{500\,bp}{3,000\,bp/minute} = \frac{1}{6}\,minute = 10\,seconds. \tag{3.3}$$

The variable T_j, with $j \in (1, n)$ specifies the replication time at every discrete time point during the simulation. An algorithm for the DNA replication has been implemented as follows. At every time point T_j the program reviews the array O to find the origins that initiate at that time. If found, the Boolean Variables for these origins in O are set to "TRUE", indicating that they have fired and cannot do so again. Furthermore, the Boolean Variables in Ar at the origins positions (e.g. Ar_{ori1} and Ar_{ori2}) are set "TRUE" as well, indicating that these regions now have been replicated. For simplicity, the activation of origins is assumed to occur at the beginning of the time steps, for which reason a unit is either replicated completely or not at all. The discretization error introduced by this approximation decreases with the DNA unit size. Every origin issues two replication forks upon activation, each traveling in opposite directions in the course of the chromosomal duplication. Therefore, at time point T_{j+1} the program checks if the positions left and right of a replicated region (e.g. Ar_{ori1-1}, Ar_{ori1+1} and Ar_{ori2-1}, Ar_{ori2+1}) have not been replicated (set "FALSE") yet, and if so, sets the Boolean Variable to "TRUE". In this manner the replication forks migrate in both directions, until they meet either the end of the chromosome, or a region that has already been replicated. Every position of every replication fork is stored at every time point of the simulation. The way of every replication fork through the genome during the simulation can be retraced and their final positions and times can be observed. The simulation stops once the whole chromosome is replicated.

3.2.3 Replication Profile Data

The spatiotemporal organization of the DNA replication process can be visualized by means of replication profiles. As schematically shown in Figure 3.2, a replication profile is the plot of the replication time as a function of the position in the chromosome. In the profile peaks correspond to origins of replication and valleys correspond to termination zones. The earlier an origin fires, the taller is its respective peak within the profile. Shoulders along the lines connecting peaks and valleys can either result from timely

collisions of a firing origin and an oncoming replication fork, or they could also be the result of change in the fork migration rate, or inefficient origins. The slope of the line connecting a peak and a valley gives the direction and rate of the fork migration.

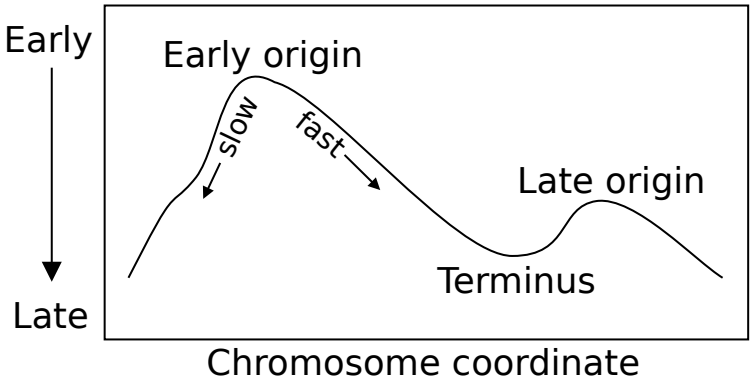


Figure 3.2: **Schematic representation of a replication profile.** A replication profile shows the replication time as a function of chromosomal position. Peaks indicate replication origins and valleys termination zones. The taller the peak in the profile, the earlier an origin fires. The slopes of the lines connecting peaks and valleys give the directions and rates of the migrating forks. This figure is adapted from Raghuraman et al. (2001).

Experimental replication profiles, which can be found in the literature (Raghuraman et al., 2001) are used to assess the model performance. The profiles are derived from a microarray based HL timing study. After growth in an isotopically dense culture medium, cells are released into S phase (after α-factor-induced G_1 phase arrest) and replicated (HL) DNAs and unreplicated [heavy:heavy (HH)] DNAs are isolated from samples collected at 10, 14, 19, 25, 33, 44 and 60 minutes (Raghuraman et al., 2001). Replication profiles for all chromosomes can be found in the Appendix B, Figure B6, where the original data from Raghuraman et al. (2001) were used to recalculate experimental replication profiles. Figure 3.3 shows the replication profile for chromosome II as a showcase. Furthermore, the data was used to calculate the total replication time for all chromosomes. Subtraction of the highest peak from the lowest valley yields the total replication time.

3.2.4 Software

The spatiotemporal model has been implemented and analyzed with the programming language *Python* (van Rossum, 1995) and the *R* statistics environment (R Development Core Team, 2007).

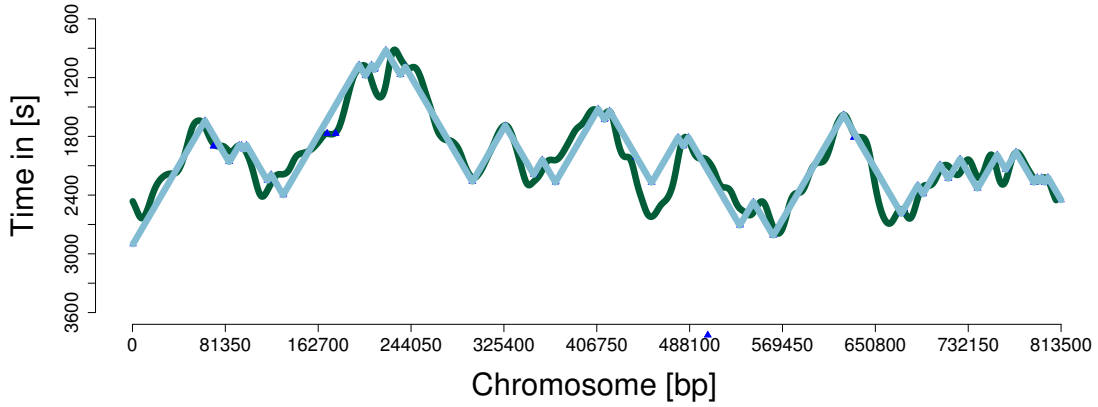


Figure 3.3: **Replication profiles of chromosome II.** The green curve is recalculated according to the microarray-based heavy:light data from Raghuraman et al. (2001), whereas the blue one represents the simulated profile obtained with the computer model. The replication time in seconds is plotted as a function of chromosome coordinate in base pairs (bp).

3.3 Results

3.3.1 Generation of Replication Profiles

The simulation of the chromosomal duplication has been performed, as described in section 3.2 with a fork rate value equal to 3 kb/minute. Sixteen replication profiles were generated, one for each chromosome, in order to highlight the spatiotemporal organization of the simulated DNA replication. Figure 3.3 shows the replication profiles for chromosome II. The smooth curve is recalculated from the data provided by Raghuraman et al. (2001), as described in section 3.2 and the straight curve shows the simulated profile. All essential features of the experimental profile were captured in the simulation.

However, we observed a deviation in the slope of the lines, representing the speed of the fork migration. The lines of the simulated curve are straight, for a constant migration rate is implemented, whereas the experimental curve is smooth with a varying slope, indicating different fork rates. Most simulated regions reflect experimental data with high accuracy and only few regions with lower accuracy. We found similar results for all 16 chromosomes (see Appendix B, Fig. B6). As reported in the work of Raghuraman et al. (2001), the fork rates range from 0.5 to 11 kb/minute with a mean of 2.9 kb/minute. Changes (increase or decrease) in the value of the fork rate could lead to different results in the computed simulations, implying more precise results in some regions and less accuracy in other regions. Additionally, it is likely that, due to differential activation of inefficient origins, the direction of fork migration during DNA synthesis may change from one cell division to the next. In accordance, it has been shown in mammalian cells that the replication speed controls the choice of the initiation firing sites on the chromosome

(Courbet et al., 2008). However, we implement a constant fork rate in the model, since we aim at a simplifying parametrization for this still not well-defined process to create an accurate, yet comprehensive representation.

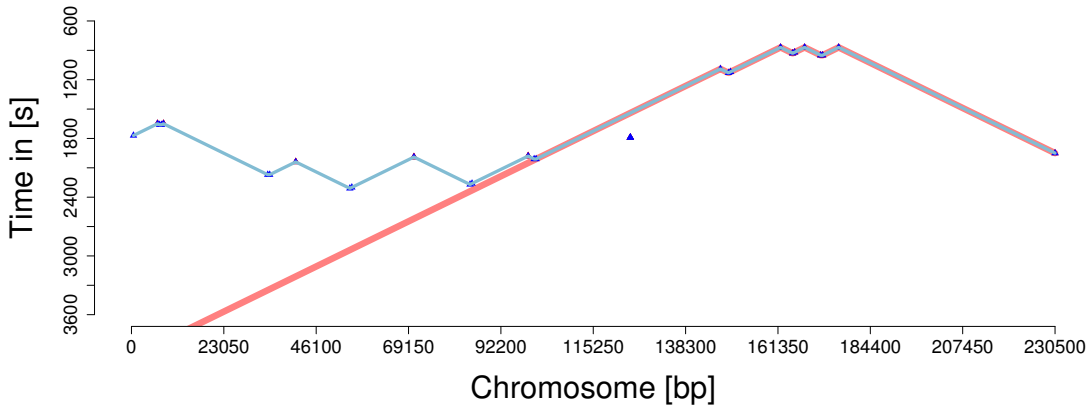
We model the chromosome duplication deterministically using the published data for locations and firing times of 454 origins of replication. Since only few data are available about origin firing efficiency (Yamashita et al., 1997), which is nonetheless known to be a key property of the origin activation, we included origin efficiencies in an implicit way. We regarded the efficiencies of a subset of all origins (454 out of 732 reported in the OriDB) as to be 100%, which is a strong assumption. However, an approximation of the replication with 454 origins that fire with an efficiency equal to 100% represents a single replication event in a cell with 732 origins that fire at about 60% average efficiency. Since the number of actively engaged origins per cell cycle has been reported to be roughly around 400 (Wyrick et al., 2001; Takeda and Dutta, 2005), this approximation seems reasonable. Employing this approach, the model does not represent a single cell behavior *per se* (no intrinsic noise in efficiencies and firing times) but reflects the average of a cell population. In other words, the model stands for a likely replication event in the average single cell, because it has been parametrized with population averaged data.

3.3.2 Chromosome Duplication in a *clb5*△ Mutant

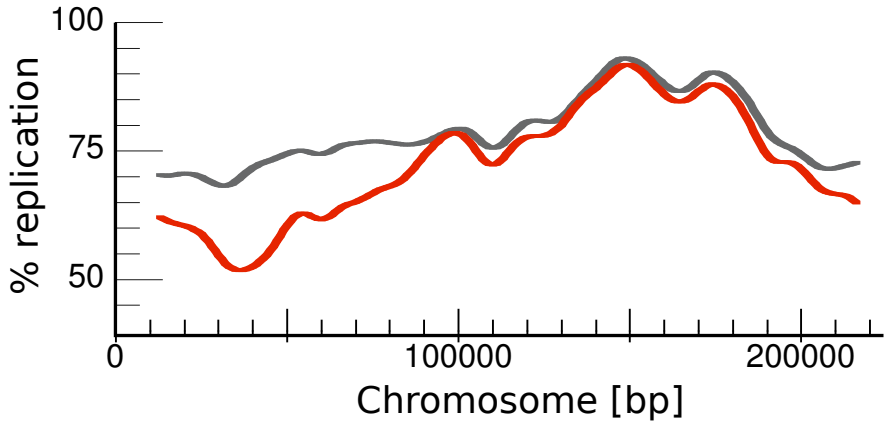
The activation of the replication machinery has still to be highlighted in many of its regulatory events, but a relevant step is the phosphorylation of different substrates by the Cdk1-Clb5/6 kinase complex that induces the firing of the DNA replication origins (Bell and Dutta, 2002; Takeda and Dutta, 2005). Barberis et al. (2007) described the steps which lead to the firing of DNA replication origins with a simple probabilistic model that considers the availability of the Cdk1-Clb5/6 nuclear concentration as main input. The model offers an explanation for the replication status of specific mutants which influence the entry into S phase, emphasizing a correlation between Cdk1 activity and timely origin activation. Along theses lines, *clb5*△ cells suffer a significant decrease in the firing efficiency of some origins, in particular for those classified as late S phase origins (Donaldson et al., 1998).

In the work of McCune et al. (2008), the activation of the replication origins has been investigated. They analyze cells that lack one initiator factor of DNA replication: Clb5. Therefore, we tested the model in the *clb5*△ mutant. Operatively, we stopped origin firing at 1,645 seconds, because it represents the mean value of the distribution of the experimentally determined origin activation times (see Appendix B, Fig. B5). Computed and experimental replication profiles for chromosome I in a *clb5*△ mutant are reported in Figure 3.4 (McCune et al., 2008). We found that multiple zones suffer significant delays in replication, whilst others are unaffected. Interestingly, the delayed regions correspond to CDRs. The CDRs match sequences of the genome which on average replicate late in S phase (Alvino et al., 2007; Raghuraman et al., 2001) and each of the late replication origins reported in the work of Donaldson et al. (1998) resides in CDR regions. The simulated replication profiles for all chromosomes in the *clb5*△ mutant environment are reported in the Appendix B, Figure B7. Figure 3.5 (a) summarizes

the results and for comparison shows the experimentally determined CDR regions from McCune et al. (2008) as well (Fig. 3.5 (b)). In detail we found a perfect match for nine chromosomes (from I to VIII and XI), a good fit in the majority of the sequence length for chromosomes IX, X and XIV and a small or no match for chromosomes XII, XIII, XV and XVI.



(a)



(b)

Figure 3.4: **Simulated (a) and experimental (b) replication profiles for chromosome I in a *clb5*△ background.** (a) The blue line represents the simulated wild type profile and the orange one represents the computed profile for the *clb5*△ mutant. (b) The wild type profile is shown in gray and the profile for the *clb5*△ mutant is shown in red. Figure (b) is adapted from McCune et al. (2008).

The analysis is in agreement with the fact that the *clb5*△ mutant only affects late origins, whereas the early origins fire normally. Therefore, the precise time at which origins stop to fire in absence of Clb5 is important. We use 1,645 seconds as the time

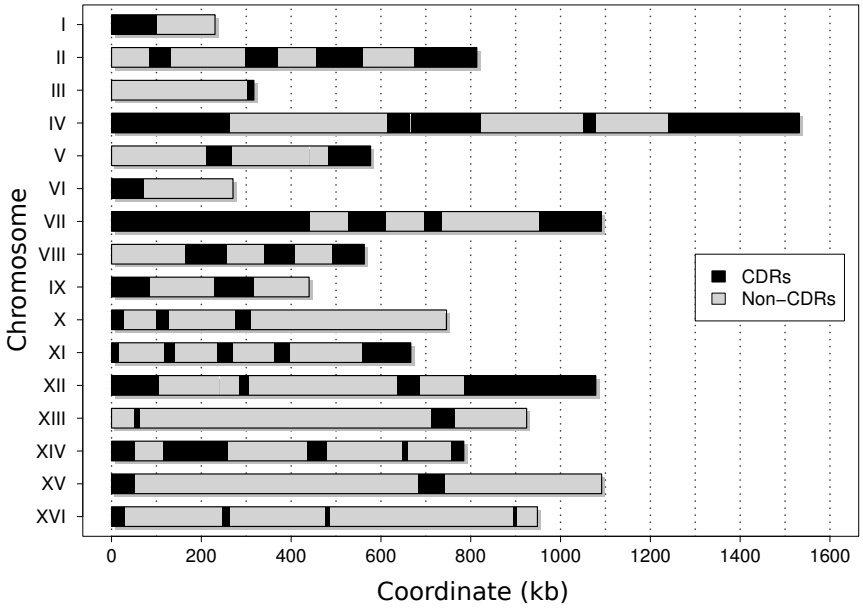
point after which there is no more origin activation. Thus, the origins are divided in an early half (Clb5-unaffected) and in a late half (Clb5-affected). However, it is likely that Clb5 activates every origin not at the same time in every cell cycle, but with a certain variation. Intrinsic noise will affect the time of the activation of the Clb5-dependent origins that will resemble more likely a time span (of some seconds or minutes). Therefore, the considered value of 1,645 is an approximation, which for some chromosomes might be quite accurate, but for others it might not be. This affects the results we observed in the following way: the chromosomes containing more early origins will be less sensitive to *CLB5* deletion, whereas the chromosomes with more late origins will be more sensitive.

The overall agreement of the replication kinetics between wild type and *clb5*$\triangle$ in the computed and experimental profiles supports the idea of a temporal program of the origin activation in budding yeast, as predicted (McCune et al., 2008). Generally, the idea that origin sequences have evolved to be exclusively sensible for either Cdk1-Clb5 or Cdk1-Clb6 is highly intriguing. We have therefore, conducted a genetic investigation of the genomic area around replication origins to study differential properties of early and late replication origins and their sequential evolution (see chapter 5 or Spiesser and Klipp (2010)).

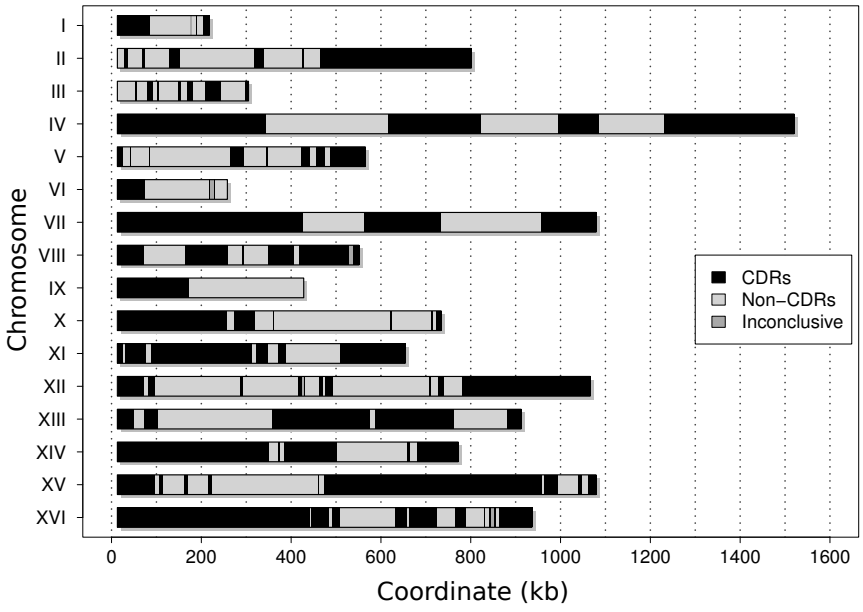
3.3.3 Impact of Origin Deletion on DNA Replication

S. cerevisiae has well-defined, site-specific origins, many of which are efficient and fire in as many as 90% of the S phases (Fangman and Brewer, 1991; Newlon et al., 1991). These characteristics lead to nearly homogeneous replication kinetics (Raghuraman et al., 2001). Despite the fact that DNA replication in budding yeast seems to follow a temporal program of origin activation, it has been reported that there is a stochastic component which can influence the process (Czajkowsky et al., 2008; McCune et al., 2008). In fact, the activation of some origins in the CDR regions more closely fits a disordered, stochastic firing pattern. They show no peak time of firing or are activated over a broad distribution of activation times in different cells in the population (McCune et al., 2008). In addition, it has been reported that variants of a stochastic firing model are compatible with a temporal staggered initiation of the replication origins in fission yeast (Rhind, 2006; Lygeros et al., 2008). In line with this, a more generalized concept has also been proposed for budding yeast recently, where origin initiation times rather correspond to origin initiation probabilities (Rhind et al., 2010). However, experimental probability distributions for the origins are not available, yet.

In order to investigate the impact of change in the origin activation pattern on the replication dynamics, replication kinetics for all chromosomes have been computed repeatedly (30 times) with reduced sets of considered origins. The subsets are composed by random deletion of 50% of the original origins. This accounts for the change in environmental conditions (i.e. stress condition, checkpoint activation) or inefficient firing, which could reduce the global origin firing efficiency from 60 to 30%.



(a)



(b)

Figure 3.5: **Clb5-dependent regions (CDRs) in the budding yeast genome** as determined through simulations (a) and experimentally (b) (McCune et al., 2008). CDRs are displayed in black and Non-CDRs in gray as in Barberis et al. (2010).

Comparison of the replication kinetics for chromosome II exhibited under wild type (Fig. 3.6 (a)) and perturbed (Fig. 3.6 (b)) conditions shows that a 50% deletion of replication origins yields a prolonged chromosomal replication time. However, we do not observe fundamental alterations in the general shape of the replication kinetics, which indicates that conditional change leading to a 50% efficiency reduction of origin firing does not change the replication dynamics of the chromosomal duplication.

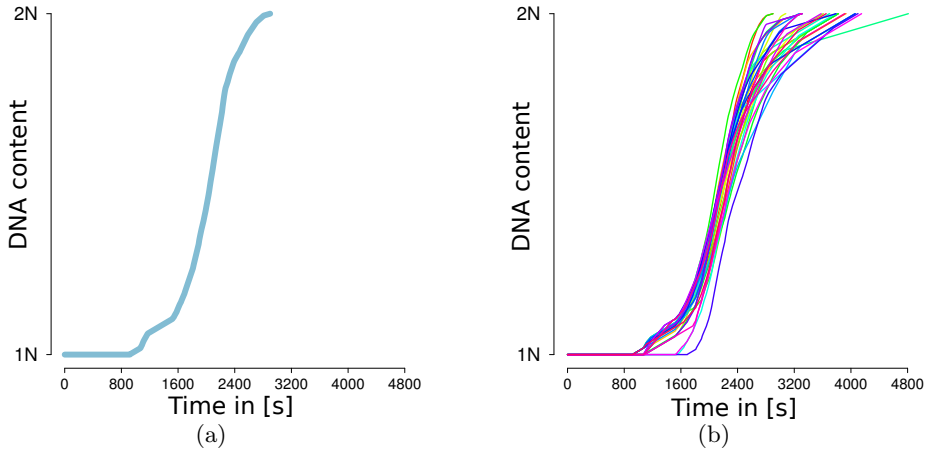


Figure 3.6: **Simulated replication kinetics of chromosome II.** The simulations are performed for wild type (a) and perturbed conditions (b). In the case of perturbed conditions, the simulation has been performed considering 30 reduced sets of replication origins derived from the random deletion of 50% of the original origins.

Moreover, we found that for most chromosomes the replication kinetics seem to show a remarkable resistance to origin reduction (see Appendix B, Figs. B8, B9). The chromosomal duplication initiates within a short timeframe, which is consistent throughout the replication process and only disperses towards replication termination. Concerning retardation, we found that 50% of origin deletion leads on average to a $\sim$ 12 minutes delay in duplication completion for chromosome II, as compared to the wild type. The remaining chromosome kinetics show similar results (see Appendix B, Figs. B8, B9). The outcome of the random perturbation of the system shows that the replication process is robust against firing failure or efficiency variation and suggests that the replication kinetics displayed by a cell can be widely independent from the temporal program of the origin activation.

3.3.4 Simulating a Stepwise Loss of Origin Function

Despite the contribution that multiple origins per chromosome may make to efficient genome duplication in *S. cerevisiae*, it is widely accepted that there are more replication origins than needed for the timely replication during the S phase (Bielinsky, 2003). In fact, several origins on chromosome III can be deleted without substantially affecting

the ability to faithfully inherit this chromosome during cell division (Dershowitz and Newlon, 1993; Dershowitz et al., 2007).

To further understand the relationship between origin activation and replication time, we simulated the chromosomal replication with a decreasing number of active origins and monitored the change of the replication time. In the previous simulations we have observed that during perturbation of the system, the replication kinetics for the chromosomes are very similar, even though they are replicated with different sets of origins. Therefore, we ignored which specific selections of origins were used in the simulations and thus, studied the relationship between the number of activated origins and the replication time directly. To this end, we used the same chromosomal location for origins and the same firing times, only the activated origins change randomly. The model predicts how the replication time of the average replication event would change, if a certain percentage of the origins were to be defective, deleted or inefficient. It is difficult to investigate the direct effect of activated origins and replication time in living systems, because the deletion of the origins often leads to the activation of adjacent usually inefficient/dormant origins. This mechanism ensures to the cell the successful chromosomal replication. Therefore, a systematic computational study is useful to highlight the relationship between a controlled quantity of active origins and the replication time.

Mean replication times for descending percentages of active origins (from 90% to 10%) have been computed for all chromosomes. The origin sets have been reduced stepwise (10%) and randomly selected. The simulations for every fraction of remaining origins were repeated 10,000 times. Mean and standard deviation for every fraction of remaining origins are displayed for every chromosome (Fig. 3.7 and Appendix B, Fig. B10). The average delay for 50% remaining origins is summarized in Table 3.1.

The calculations for the chromosome II show that, with a decreasing percentage of remaining origins, the mean replication time increases, as well as the standard deviation (Fig. 3.7 (a)). This is the case for all chromosomes, although the intensity of the increase differs amongst the chromosomes. Interestingly, the experimentally assessed duplication times can be obtained using only a certain subset of activated origins and the subsets are different for every chromosome and composed randomly. An example is reported for chromosome XVI (Fig. 3.7 (b)). The experimental replication time, derived from Raghuraman et al. (2001), is indicated as a green line. The simulation shows that chromosome XVI duplication could be achieved, in the experimentally measured time, with subsets of only 50-60% randomly selected origins (Fig. 3.7 (b); Tab. 3.1), as indicated by the intersection of the green and the blue line. This percentage differs for every chromosome and for some chromosomes the replication can only be simulated in the appropriate time with 100% of the origins, e.g. for chromosome II (Fig. 3.7 (a); Tab. 3.1).

The simulations mirror the robustness of the replication process against perturbations in origin firing, as a result of loss of the origin function or change in the total efficiency. Using a systems study, we highlight the relationship between origin activation and replication time in the average cell population in budding yeast. The reduction in origin firing up to, e.g. 50% in chromosome II can be compensated within the system resulting in a delay of about 12 minutes in replication completion (Figs. 3.6, 3.7). This is the case

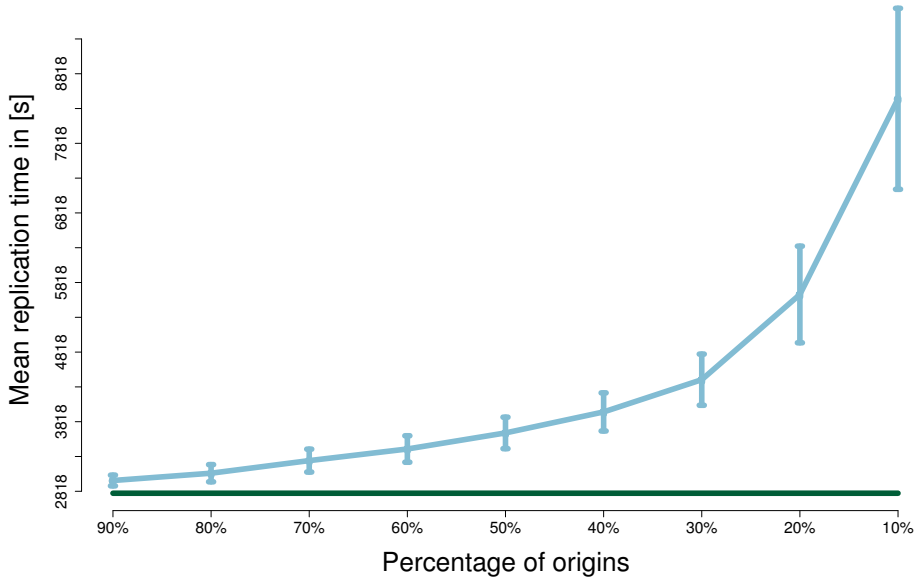
Chromosome	Average delay (50% origin deletion)	Active origins in % (crossing experiments/simulations)
I	7 min 00 sec	30 - 40
II	12 min 36 sec	100
III	2 min 29 sec	50 - 70
IV	18 min 54 sec	70 - 90
V	15 min 29 sec	90 -100
VI	3 min 52 sec	60 - 90
VII	12 min 37 sec	100
VIII	9 min 53 sec	40 - 50
IX	5 min 59 sec	30 - 40
X	11 min 22 sec	40 - 50
XI	13 min 30 sec	100
XII	14 min 34 sec	40 - 50
XIII	16 min 17 sec	50 - 70
XIV	20 min 48 sec	100
XV	17 min 08 sec	100
XVI	14 min 01 sec	50 - 60

Table 3.1: **Average delay in chromosomal duplication time**, under 50% origin deletion condition, calculated after 10,000 simulations of DNA replication. The percentage of origins is indicated, which is required to simulate the chromosomal duplication in the experimentally measured time.

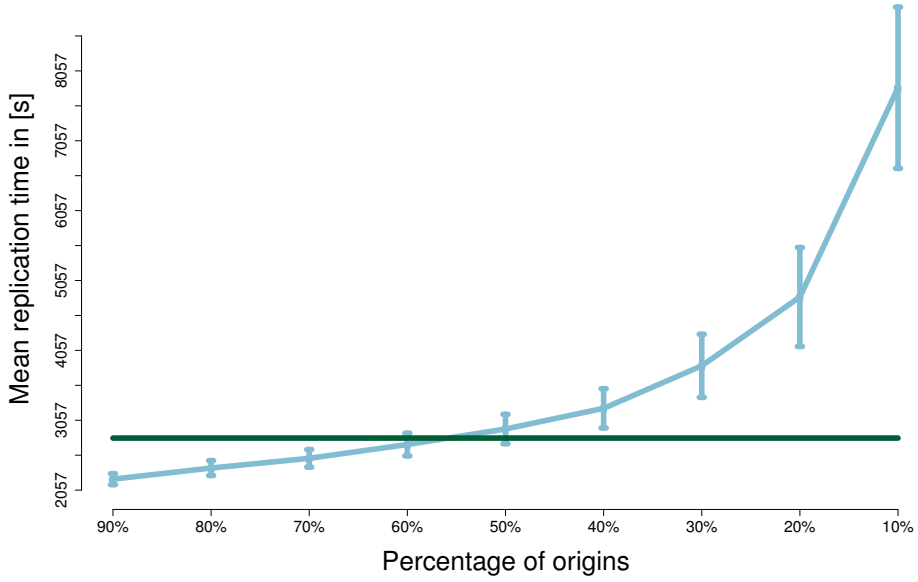
obviously only if no other late/dormant origins fire. A similar effect can be observed for the remaining chromosomes (Tab. 3.1). The average delay in chromosomal duplication increases with the size of the chromosomes (Fig. 3.8 (a)) and decreases with an increasing origin density on the chromosomes (Fig. 3.8 (b)). The origin density is the ratio between the number of origins on a chromosome and the chromosome size.

3.4 Discussion

The goal of the work, outlined in this section, was to provide a model for the DNA replication dynamics, based on replication system parameters, to study the temporal sequence of origin activation in *S. cerevisiae*. The system parameters are: (1) lengths of the chromosomes, (2) location of the origins on the chromosome, (3) firing time of the origins and (4) replication fork migration rate. The parameters used in the analysis were obtained from experimental data (see section 3.2 for details). In the spatiotemporal model of DNA replication, two limiting factors impinge the biological validity of the model: the approximation of the fork migration rate with the mean of the experimentally determined value of ~ 2.9 kb/minute (Raghuraman et al., 2001) and the implicit consideration of the origin efficiencies.

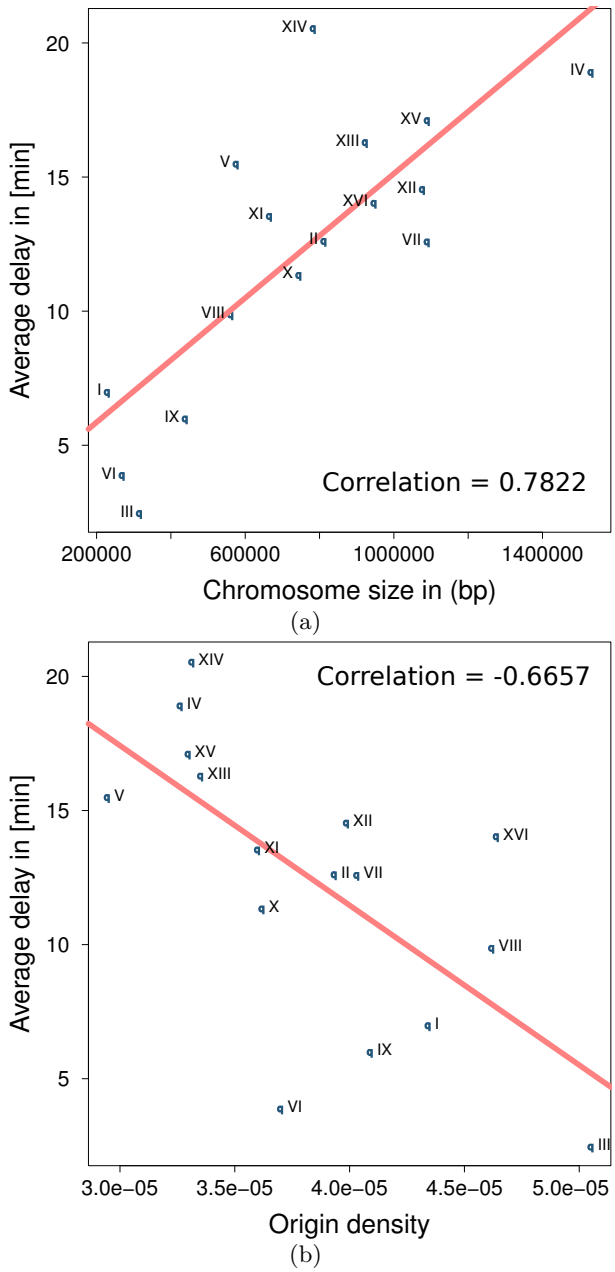


(a)



(b)

Figure 3.7: **Mean replication time (in seconds) for chromosomes II (a) and XVI (b).** Blue line represents the curve for descending percentage of the considered replication origins (from 90% to 10%). Error bars show the standard deviation of 10,000 simulations. Green lines indicate the experimental replication time, according to Raghuraman et al. (2001).



(a)

(b)

Figure 3.8: **Average delay in chromosomal duplication time (in minutes) over length (a) and origin density of the chromosomes (b).** Average delay is calculated after 10,000 simulations of DNA replication under perturbed conditions (50% origin deletion). Origin density is the ratio of number of origins on a chromosome and chromosome size. The correlation of the data and coefficient are shown in orange.

The model has been used to generate replication profiles, which plot replication time as function of the chromosome coordinate. They have been compared to the replication profiles reported in the literature (Raghuraman et al., 2001). The comparison has shown that the model is generally able to reproduce the experimental replication profiles (Fig. 3.3). Some disagreements between simulations and experiments can be observed, which is essentially due to two different reasons.

First, we introduce a bias by using a single, approximated value for the fork migration rate, which means that the rates of motion in the model are constant and do not take changes of speed into account. The result is inaccuracies in the simulations of the replication profiles. The accuracy of the model could perhaps be increased by consideration of a dynamic fork rate function. Different fork rates at different chromosome regions could have either regulatory functions or could be caused by higher order structures of the chromosome (protein binding, 3-D effects, etc.). Therefore, a rate function that is adapted to those different, biological characteristics influencing the migration rate, could enhance the performance. We have directly tested this hypothesis using a stochastic model for the replication machinery motion. The stochastic model and the results are shown in chapter 4 or in Spiesser et al. (2010).

Second, we do not include single origin efficiencies as an adjustable parameter, because too few are currently available (Yamashita et al., 1997). This means leaving out a key property of the origins and, with it, its stochastic influence on the replication process. However, we based our modeling on the assumption that in one cell cycle there are about 400 origins that fire with 100% efficiency, when indeed there are much more origins (732) that could be potentially used. Thus, we approximated the overall efficiency of initiation in a cell with 732 origins at roughly 60%. Previous studies indicate that the excess of origins can help the cell to ensure the duplication under stressed conditions (Dershowitz and Newlon, 1993; Dershowitz et al., 2007). This means that our modeling reflects DNA replication of a particular cell cycle and - due to the parametrization of the model with population averaged data - it represents the average DNA replication event in a budding yeast cell. These assumptions could be relaxed when more experimental data will become available.

S. cerevisiae has a 13.5 mega bases genome distributed over 16 chromosomes and therefore, each single yeast chromosome is considerably smaller than the 4.6 mega bases *E. coli* genome. Yet, yeast replication origins occur on average every 20-40 kb, a hundred times more densely distributed than one would predict by comparison to the *E. coli* genome. The difference in fork migration rates may explain in part the need for multiple replication origins per eukaryotic chromosome, since DNA replication forks migrate at rates about 30 times slower in yeast compared to *E. coli*. Replication forks migrate at rates of about 3 kb/minute (yeast) compared to about 100 kb/minute (*E. coli*) (Raghuraman et al., 2001; Rivin and Fangman, 1980). The use of multiple initiation events per chromosome probably compensates for slower fork migration rates in maintaining an efficient rate of genome duplication and S phase progression in eukaryotic cells. However, based on the values discussed above, *S. cerevisiae* would need about 100 replication origins to duplicate its genome at a rate sufficient to accommodate its S phase, about four times less than the current estimates for origin numbers in this organism (Raghuraman

et al., 2001; Wyrick et al., 2001). Therefore, for the purpose of genome duplication, yeast replication origins are redundant and it is interesting to investigate the relation between the number of active origins and the replication time. We used the model to systematically study this relationship. To assess the impact of particular sets of origins on the replication time, we computed replication kinetics under wild type and perturbed conditions. The replication kinetics mirror the dynamic of the replication system and are therefore, a useful tool to investigate the influence of conditional changes on the system. Perturbing the replication process by severe loss of the replication origin function due to their random deletion showed only little influence on the replication dynamics (Fig. 3.6). Therefore, we could neglect the effect of specific origin sets on the time of DNA replication and systematically deactivate an increasing number of origins. As expected, the analysis showed that the more origins that were deactivated, the more time was needed to complete the chromosomal duplication, but interestingly highlights that the experimentally assessed duplication times can be obtained using only a certain subset of activated origins (Fig. 3.7).

In the model, we implemented directed movement for the DNA polymerase. Therefore, we do not allow backward movement during our simulations and thus, we argue that the anticipated relationship between distance and time is close to linear. However, this linear relationship is not directly visible in our results since we monitor the mean replication time with respect to the removal of origins, which one could also interpret as a system with an increasing failure rate over time. The replication time is dependent on the longest distance that a replication fork covers, which is the maximum value of the inter-origin spacing (extreme value of the distance between the origins). Successive removal of origins from the chromosome results in longer distances between the remaining origins. If we interpret this system as one with an increasing failure rate over time, we could describe this system with an extreme value distribution, being in our case the distance between the origins. However, we can only describe our results to a certain extend by such an extreme value distribution, because naturally the firing times influence the system as well. Normally distributed firing times (Appendix B, Fig. B5) lead to exponentially distributed waiting times and this effect smoothens the curve that we obtain.

The analysis showed that the replication system is robust against perturbations. This suggests that a purely deterministic program of the origin activation in budding yeast might be enough only at the first glance on the system, but possibly not to describe all of its properties. If a temporal program is influenced by stochastic patterns, we would expect the replication system to cope more easily with perturbations and therefore, to successfully complete DNA replication with hardly any substantial changes in the dynamics of the replication. Where in the purely deterministic system the defects in origin firing due to a perturbation would be more severe (stress condition, origin deletion, inactivation of some specific initiation factor which stimulate origins activation), a stochastic component would always provoke some random activation of origins. Hence, a stochastic influence can increase the robustness and thus, be advantageous for the system.

Moreover, we found that the length of a chromosome and its origin density have an impact on the robustness. In fact, the replication delay under perturbed conditions is increased for larger chromosomes, whereas the average delay is decreased for the

chromosomes that have a higher origin density (Fig. 3.8). Consequentially, the increase in the delay could be interpreted as a decrease of robustness and the decrease in the delay could be seen as an increase in the robustness. Altogether, this suggests that smaller chromosomes with higher origin density are more robust towards perturbation. It is tempting to speculate that this could be an explanation for why organisms have evolved to rather have a number of smaller chromosomes, instead of only a single large one. In any case, it seems favorable for an organism to possess a high number of origins, a selection of which is finally activated to duplicate the DNA within the required timeframe.

In conclusion, we have successfully constructed a simple, yet accurate deterministic spatiotemporal model for DNA replication in budding yeast, which reproduces the trends exhibited during chromosomal duplication. The results of our analysis suggest that the replication system is robust against perturbations and that there might be a stochastic component in the temporal activation of the replication origins, especially under perturbed conditions. The observed robustness could be tested experimentally by deleting origins progressively and evaluating the replication time for each chromosome. Our future goal would be to investigate the influence of stochasticity on the temporal program of origin activation in budding yeast more closely. Noteworthy, a partially deterministic and partially stochastic order of DNA replication was already addressed in a model for DNA replication in mammalian cells (Takahashi, 1987). In the light of this evidence, our model could well be suitable for further and more accurate investigation of the temporal origin activation in budding yeast, in particular as soon as experimental data concerning origin efficiencies will become available. Moreover, the computational analysis could be extended to eventually link DNA replication to the classical cell cycle machinery and its relevant checkpoints.

4 What Influences DNA Replication Rate in Budding Yeast?

This chapter is dedicated to a stochastic model for the motion of the DNA replicating machinery along the DNA template. The model is used to study differences in elongation times found in budding yeast. Futhermore, it is used to reassess the assumption made in chapter 3 that the migration rate of a replication fork can be described using a global average rate. The chapter is based on:

T. W. Spiesser, C. Diener, M. Barberis and E. Klipp. What Influences DNA Replication Rate in Budding Yeast? *PLoS One*, 5(4):e10203, 2010.

4.1 Introduction

DNA replication begins at specific locations called replication origins, where helicase and polymerase act in concert to unwind and process the single DNA filaments. The sites of active DNA synthesis are called replication forks. The density of initiation events is low when replication forks travel fast and is high when forks travel slowly (Goldar et al., 2008). Despite the potential involvement of epigenetic factors, transcriptional regulation and nucleotide availability, the causes of differences in replication times during DNA synthesis have not been established satisfactorily, yet. Here, we aimed at quantifying to which extent sequence properties contribute to the DNA replication time in budding yeast. We interpreted the movement of the replication machinery along the DNA template as a directed random walk, decomposing influences on DNA replication time into sequence-specific and sequence-independent components. As shown in section 4.3.1, we found that for a large part of the genome the elongation time can be well described by a global average replication rate and thus, by a single parameter. However, in section 4.3.2 we also show that there are regions within the genomic landscape of budding yeast with highly specific replication rates, which cannot be explained by global properties of the replication machinery. I show and discuss here (see section 4.4) that even beyond the level of initiation there are effects governing the replication time that can not be explained by the movement of the polymerase along the DNA template alone. This allows us to characterize genomic regions with significantly altered elongation characteristics, independent of initiation times or sequence composition.

4.2 Materials and Methods

4.2.1 Model Formulation and Assumptions

The general assumption of this work is that observed replication rates, that can be found in literature, are governed by two different and independent aspects, one that is sequence-specific and one that is not. It is the combination of both aspects that determines the shape of the experimental replication profiles (Raghuraman et al., 2001) and the dynamics of DNA replication. However, it is currently not known to which extent both factors contribute to the observed dynamics, nor whether these contributions are locally restricted or not. There are global properties influencing the replication rate (like the nucleotide composition), as well as e.g. histone acetylation/methylation or active transcription, which vary throughout the genome and are therefore, local quantities. We assumed that the replication time of the profiles (T_{prof}) is composed of the following: the time that the replication machinery needs in terms of reaction kinetics (nucleotide incorporation) and motion (T_{seq}), the time that is needed to account for active transcription or any other local regulation (T_{reg}) and an error (ϵ) standing for random fluctuations, thus: $T_{prof} = T_{seq} + T_{reg} + \epsilon$. This equation also exemplifies our approach: we decomposed the experimental data (T_{prof}) into the different components. We did so, by describing and therefore capturing the underlying, seizable part of the system (T_{seq}) filtering it from the data, to unravel the error (ϵ) and the unknown, regulatory component (T_{reg}) of the data.

Genomic sequences for all the 16 chromosomes of budding yeast were obtained from the NCBI reference sequences database (Pruitt et al., 2007). Information about the replication dynamics in budding yeast was extracted from recently published whole genome replication profiles (Raghuraman et al., 2001). A detailed description of the replication profiles can be found in section 3.2. As an example, the profile for chromosome II is shown in Figure 4.1. The slope of the line connecting an origin (peak) and a termination zone (valley) shows the direction and the rate of the fork migration. Replication profiles represent an average of population and not single cell data and therefore, caution must be taken in directly relating those profiles to the elongation time of the individual replication forks. Raghuraman et al. (2001) calculated the profiles as means over several individual measurements. Therefore, we can not expect to characterize the level of variation within the data and thus, the inherent stochasticity. However, it is possible to calculate the mean value of the stochastic process that governs the replication dynamics. Additionally, profiles obtained from the literature have been smoothened prior to publication and thus, been transformed to a continuous curve where the original peaks and valleys of the profile at the replication origins are flattened. This leads to a slight distortion of the data.

We approximated the maximum error this effect imposed on the replication profiles. This error can be quantified by measuring the lengths of chromosomal regions within the profile that show a non-zero curvature, thus $\left|\frac{d^2y}{dt^2}\right| > 0$. Multiplying the lengths of those regions, **L** (in base pairs), by the inverse of the average overall replication rate,

α^{-1} (in seconds per base pairs), yields the error distribution

$$\boldsymbol{\epsilon}_{curv} := \mathbf{L} \cdot \alpha^{-1}. \tag{4.1}$$

Furthermore, the profiles contain the combined information of the initiation (or firing) time of the origins and the time required for the elongation for every chromosomal region. In this chapter we shall refer to the genomic sequence between one peak and one valley in the profile as a "segment". For those segments we calculated the elongation time as the time difference between the corresponding peak and valley (as shown in Fig. 4.1).

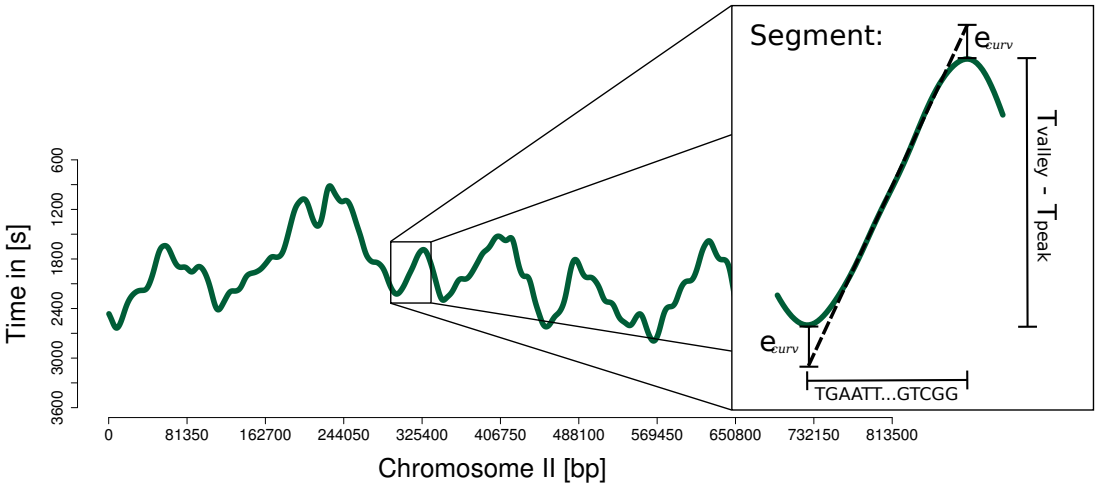


Figure 4.1: **Schematic view of the data processing procedure.** The genomic sequence between one peak and one valley in the experimental profiles (Chromosome II is shown as an example (Raghuraman et al., 2001)) is called "segment". We calculated the elongation time as the time difference between the corresponding peak and valley, where ϵ_{curv} denotes the error caused by data smoothing.

Thus, a single segment s^i is assigned to a single elongation time T^i_{prof} which we decomposed into

$$T^i_{prof} = T^i_{seq} + T^i_{reg} + \epsilon^i. \tag{4.2}$$

For T^i_{seq} we allowed a direct dependence on the nucleotide composition of the sequence, which is the frequency of each nucleotide within the segment. The remainder consists of a normal-distributed error term $\epsilon^i \propto \mathcal{N}(\mu, \sigma)$, as defined in equation 1.23, and a specific time T^i_{reg}. T^i_{reg} denotes some unknown local influence on the replication time and does not follow the normal distribution of the error. We allowed a non-zero mean (μ) here since we might have systematic global errors. For example $\boldsymbol{\epsilon}_{curv}$ is also contained in $\boldsymbol{\epsilon}$. This directly imposed a statistical test for identifying segments with a non-zero T^i_{reg} by comparing against the null-hypothesis of the error distribution of the ϵ^i. To this end, we filtered the individual T_{seq} from the elongation times T_{prof} by building a mathematical

model which specifically describes T_{seq}.

Here, we assumed that the replication machinery movement on the DNA segment follows a directed random walk, where the probabilities for the movement and the corresponding waiting and step times were only dependent on the current position (base) of the replication machinery and independent of the previous or next position. Furthermore, since the data of Raghuraman et al. (2001) only indicate the movement of the replication machinery and does not give detailed information about leading and lagging strand polymerization, we made further assumptions. The following components are not modeled explicitly but assumed as part of the replication machinery: helicase Mcm2-7 with associated factors, polymerases δ and ϵ, polymerase α-primase and ligase. We further assumed that the synthesis of the leading and the lagging strand occurs in parallel.

For the movement we assumed that the replication machinery would either moves forward with a base-dependent probability $p(X)$ for base X or wait with probability $1-p(X)$ ($X \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$). For a finite sequence this yields a total step number $N_{tot}(X)$ for each base being the sum of forward (f) and waiting (w) steps ($N_f(X) + N_w(X)$). Here, the forward step would take a characteristic time $t(X)$ and the waiting step a time $w(X)$ (illustrated in Fig. 1.4). Due to the spatial independence the probability for k forward steps for base X now follows a binomial distribution, as defined in equation 1.22, thus

$$\mathbf{P}(k, X) = \binom{N_{tot}(X)}{k} p(X)^k (1 - p(X))^{N_{tot}(X)-k} \tag{4.3}$$

with expected forward steps

$$N_f(X) = E(k, X) = N_{tot}(X)p(X), \tag{4.4}$$

where $E(k, X)$ denotes the expectation of the binomial distribution, as defined in equation 1.18. However, since $N_{tot}(X) = N_f(X) + N_w(X)$ and $N_f(X)$ being the (expected) number of forward steps for base X, we can derive the expected number of waiting steps by the number of forward steps, since

$$N_{tot}(X)p(X) = N_f(X) \tag{4.5}$$
$$(N_f(X) + N_w(X))\, p(X) = N_f(X) \tag{4.6}$$
$$N_w(X)p(X) = N_f(X)(1 - p(X)) \tag{4.7}$$
$$N_w(X) = N_f(X)(p(X)^{-1} - 1). \tag{4.8}$$

This formulation is important since the information obtained from the profiles is the number of forward steps for each of the bases (simply the base counts in the segment). Thus, receiving the number of forward steps for each base $N_f(X)$ from the segment lengths we could derive the expected replication time as the sum of times required for

each subset of bases,

$$\hat{t} = \sum_X N_f(X) \left(t(X) + \left(p(X)^{-1} - 1 \right) w(X) \right). \tag{4.9}$$

Defining the column vectors $\mathbf{p} = (p_X)^T$, $\mathbf{t} = (t_X)^T$ and $\mathbf{w} = (w_X)^T$ and setting $\mathbf{N}$ to be the $(F \times 4)$ matrix with the base counts for each of the F segments in its columns, we can concisely derive the segment-depending replication times via

$$\hat{T} = \mathbf{N} \left(\mathbf{t} + \left(\mathrm{diag}(\mathbf{p})^{-1} - \mathbf{1} \right) \mathbf{w} \right), \tag{4.10}$$

where $\mathbf{1}$ is the Identity matrix. Equation 4.10 is, under the given assumptions, the most general description of the time required for the replication of a single segment. We call it here *model 1*. It is the most complex model because it allows different parameters for each of the four bases (12 parameters in total).

However, one may also make further assumptions in order to reduce the complexity of the model and test whether the four bases have the same influence. In this special case, where we assumed independence of the base itself, the matrix $\mathbf{N}$ becomes a column vector where each row entry denotes the length of the segment and the parameter vectors become scalar. The approximated replication times are then given by

$$\hat{T} = \mathbf{N} \left(t + (p^{-1} - 1) w \right). \tag{4.11}$$

The description in equation 4.11 is called *model 2*. It uses the same parameters for each of the four bases (3 parameters in total).

Finally, we further simplified the model to a version where the second term was summarized into a single parameter $\tilde{t} = t - (p^{-1} - 1) \cdot w$, yielding a completely linear model of the form

$$\hat{T} = \mathbf{N} \cdot t. \tag{4.12}$$

Equation 4.12 is the most simple description, called *model 3*: an average replication time per base multiplied with the length of the segment.

All filtering has been done with the most detailed description we derived (*model 1*). The other two models were solely used for model comparison.

4.2.2 Model Fitting

Models 1 and *2* were fitted to the experimental data (Raghuraman et al., 2001) by an initial global regression step followed by a local refinement step. The global step was performed using Simulated Annealing with a modified sampling step, where we used a kernel of truncated normal distributions in order to include boundaries for the parameters (all parameters were assumed to lie within [1e-8, 1]) (Kirkpatrick et al., 1983). The local refinement step was executed using the L-BFGS-B algorithm with the same boundaries (Zhu et al., 1997).

As a goal function we chose the residual sum of squares ($RSS = \boldsymbol{\epsilon}^2$ as defined in

equation 1.31) given by the measured values T and the approximated values $\hat{T}$, thus

$$RSS = (T - \hat{T})^T (T - \hat{T}). \tag{4.13}$$

The regression was performed for 1000 uniformly distributed initial values (in the range [1e-8, 1]) for the parameters which enabled us to derive the parameter correlations. The remaining replication times, or filtered times, were then calculated as the difference of the experimental and the mean (see equation 1.18) of the fitted replication times

$$T - \frac{1}{1000} \sum_{i=1}^{1000} \hat{T}_i \tag{4.14}$$

and their distribution and remaining correlation to the segment lengths was computed. For all correlation measures, we used the Spearman rank correlation, as defined in equation 1.25.

In order to quantify the effects independent of the underlying sequence or segment length the filtered times were first approximated by a normal distribution. The rationale behind is that a normal distribution would indicate a combination of random processes being responsible for the residuals whereas all deviations from that distribution would indicate some form of regulation (see section 1.4.2 and De Moivre (1738) for details on approximating the distribution of a combination of random processes). The parameters for the normal distribution were approximated by robust measures, namely the median for the mean (see section 1.4.2) and the median absolute deviation, as defined in equation 1.21 for the standard deviation.

In a second step we identified all segments whose remaining replication time (deviation from the approximated segment-dependent replication time) was significantly different from the prior normal distribution on a significance level of 0.05 with the Holm-Bonferroni correction applied (Holm, 1979). This also ensured that the smallest significant remaining replication time was still larger than the largest error which we can expect due to the smoothening of the profiles. Thus, the significance can not be explained by the data smoothening.

4.2.3 Model Ranking

In a last step we ranked the models according to the Akaike Information Criterion (AIC), as defined in section 1.4.3 (Akaike, 1974). The AIC is a tool for model selection, which means it can be used to compare competing models with one another. Here, the AIC has been calculated on the basis of two different statistical measures defined in equations 1.33 and 1.34 with n equal to the number of observations, RSS as outlined in equation 4.13 and with

$$R^2 = 1 - \frac{(T - \hat{T})^T (T - \hat{T})}{(T - \overline{T})^T (T - \overline{T})} \tag{4.15}$$

where $\overline{T}$ is equal to the mean of T.

4.2.4 Software

All tasks were implemented and analyzed with the *R* statistics environment (R Development Core Team, 2007).

4.3 Results

4.3.1 Elongation Times are Directly Related to the Segment Lengths for a Large Part of the Genome

On the assumption that the observed replication profiles can be decomposed into a sequence-related part and a non-related part (see sections 1.2.3 and 4.2), we built a stochastic model for the replication machinery movement to characterize the first part of the equation $T_{seq} + T_{reg} + \epsilon = T_{prof}$. Therefore, the model must be able to capture the two different attributes of T_{seq} that matter the most: (1) differences in base composition of the DNA and (2) differences in lengths of segments.

We found a large dependency between the segment lengths and the experimental replication times (correlation coefficient ~ 0.82 (Appendix C, Fig. C11)). On the contrary, we found almost no dependency between the replication times and the base composition of the segments. The correlation matrix for the 12 parameters of *model 1*, calculated as described in section 4.2, shows that they are correlated in a block-like manner (Fig. 4.2 (a)). The blocks represent probabilities for the movement of the replication machinery, the transition times and the waiting times. All probabilities for the single nucleotides are slightly positively correlated to the transition times (white and light orange ovals) and negatively correlated to the respective waiting times (light blue and blue ovals). A small negative correlation between transition and waiting times (orange and light blue ovals) is observed, however, the intensity of the correlations differs amongst them. Nevertheless, we notice that the higher the chance that the replication machinery moves across a certain nucleotide, the shorter are waiting times in case the polymerase stalls (Appendix C, Figs. C12 and C13).

Figure 4.2 (b) shows a similar, yet inverse trend for the 3 parameters of the small model. The transition probability is highly positively correlated to the waiting time (orange oval). The transition time is, if so at all, slightly positively correlated to the transition probability (light orange oval) and slightly negatively correlated to the waiting time (violet oval). In other words, the higher the chance that the polymerase moves at all, the longer it waits in case of stalling.

Figure 4.2 (c) shows the filtered times for the three models. Even though the models differ in the number of parameters, *model 1* cannot describe the experimental data more accurately than the smaller *model 2* or even the linear *model 3*. Despite the difference in degrees of freedom, the residual sum of squares is only slightly smaller (0.05%) for *model 1* compared to the small and the linear ones (Table 4.1). Model ranking yields that relative to the different number of parameters the linear *model 3* performs best, the small *model 2* second best and *model 1* worst.

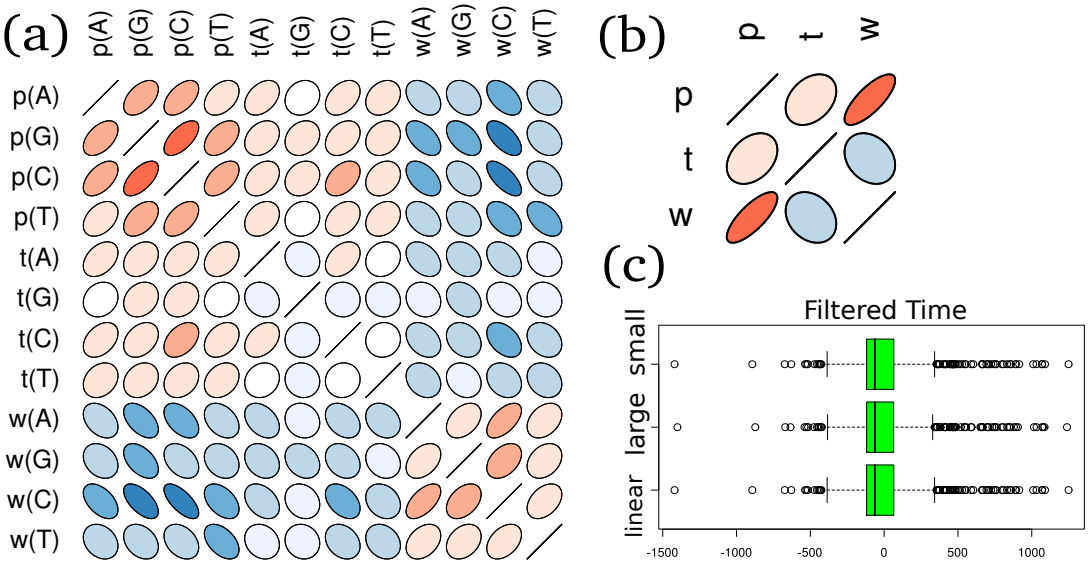


Figure 4.2: **Model comparison.** (a) Correlation matrix for *model 1*. The shape of the ellipses correspond to 95% confidence regions of a Gaussian kernel with the given correlation, as such the longer diameter of the ellipses specifies the direction of correlation whereas the smaller diameter describes how the data deviates from the line of correlation. Orange and blue colors indicate positive and negative correlations, respectively. (b) Correlation matrix for *model 2*. (c) Filtered times for the three models.

The detailed model does not fit the experimental data significantly more accurate than the smaller or the linear model. This indicates that the effect which determines the velocity of the replication machinery is largely independent of the composition of the sequence that is to be replicated. If there are differences in transition probabilities, transition times or waiting times between the nucleotides, their contribution is too small to finally determine replication rate deviations. This also holds for nucleotide pairs and triplets (Appendix C, Figs. C14, C15 and C16). Thus, apparent deviations in the replication rate cannot be explained by differences in the sequence composition.

Furthermore, despite the huge amount of experimental data points, *model 1* as well as *model 2* seem to be over-determined (see section 1.4.3); too many parameters show correlation, which indicates that one parameter can be enough to characterize the replication rate in budding yeast, as we proposed in chapter 3 and Spiesser et al. (2009).

Since base composition does not seem to play a major role and in order to test how much of the length specific correlation is captured by the model, we calculated the correlation coefficient for the filtered times and the segment lengths (Fig. 4.3). This value was significantly smaller (~ 0.05), which indicates that there is hardly any correlation left between the length of the replicated segment and the rate at which it is replicated.

	Model 1 (Large)	*Model 2* (Small)	*Model 3* (Linear)
RSS	42682347	42701178	42701178
R^2	0.535819751983114	0.535614953567048	0.535614953574985
AIC	7425.48716598	7407.78070659	7403.78070659
AIC_{R^2}	16.7267336032	-1.27282528956	-5.27282528958
Rank	3	2	1

Table 4.1: **Model statistics and ranking.** Residual Sum of Square (RSS), Coefficient of determination (R^2), general Akaike Information Criterion (AIC), Akaike Information Criterion based upon the Coefficient of determination (AIC_{R^2}) and the model rank are shown.

In conclusion, we succesfully filtered out $\sim 95\%$ of all sequence-specific rate deviations (T_{seq}) from the experimental data (T_{prof}).

4.3.2 Regions with Strongly Altered Elongation Distinctly Map onto the Budding Yeast Genome

The remaining component of the data is now ϵ and T_{reg}, which can be observed in Figure 4.3. We found that our model (*model 1*, average of 1000 different parameter sets), indicated by the median of the filtered time histogram, is slightly too slow (median = 62.7735 seconds). However, on a time scale of up to 1500 seconds, this is an error of only $\sim 4\%$. Furthermore, we observe a lower and an upper tail of the filtered time distribution, which are prominently placed outside the overlying normal distribution. These tails indicate DNA segments where the model predicts much faster or slower replication than observed in the experiments. The upper tail is more prominent compared to the lower one. However, it seems that, since the times are already filtered, in both regions other mechanisms, different from segment composition or length, influence the rate of DNA replication.

We visualized all regions of replication rate deviation for the 16 chromosomes of budding yeast (Fig. 4.4). The chromosomal regions that replicate faster in the experimental data compared to the predictions of *model 1* are shown in blue, whereas the regions that replicate slower are shown in green. The magnitude of the deviation is indicated by the intensity of the colors.

We found that only few regions replicate faster (blue), whereas many regions show significant delays in DNA replication (green). In particular, we found that only two regions on chromosome IX, one region on chromosomes XI and XII, respectively and three regions on chromosome XIV replicate significantly faster. On the opposite, the regions where replication is delayed are more frequent and scattered over nearly all chromosomes (except for chromosomes II, XIV and XV). No significant deviations could be detected only for chromosomes II and XV. The exact landscape of the filtered times and the original profiles from Raghuraman et al. (2001) for all 16 chromosomes can be found in Appendix C, Figure C17. We did not observe that regions with strongly altered elongation correlate with late or early firing origins.

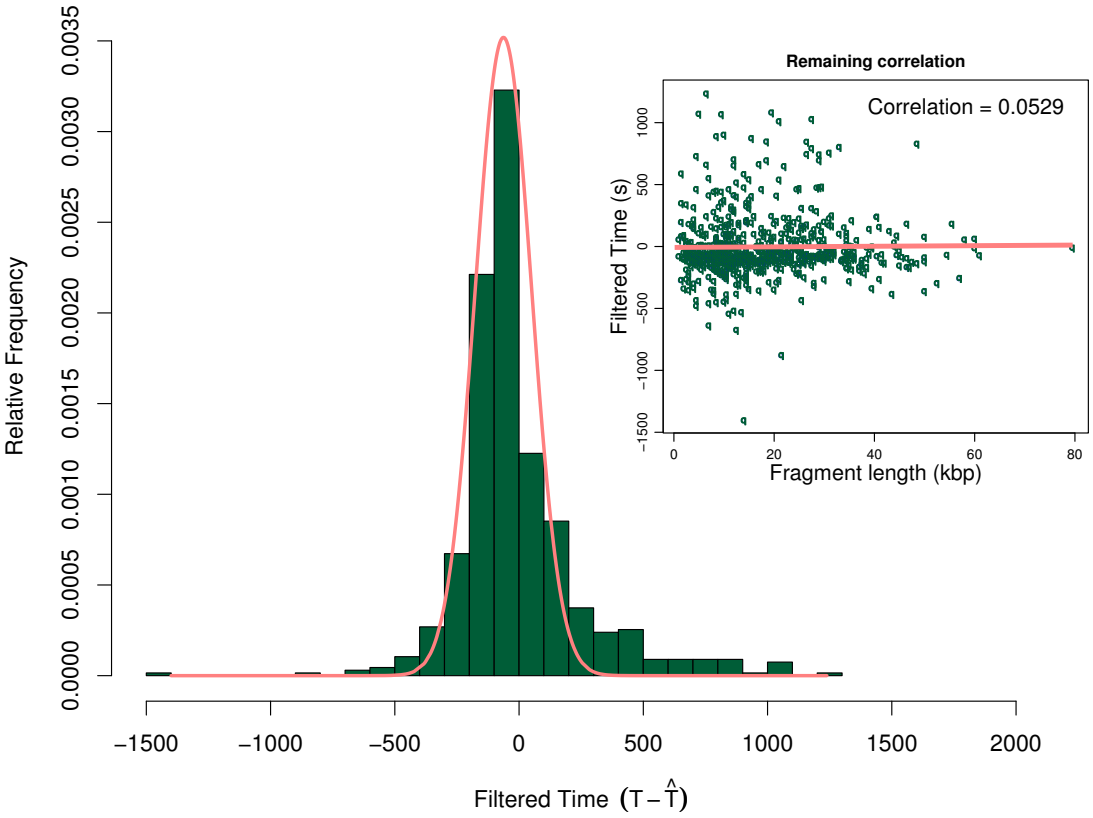


Figure 4.3: **Histogram of the filtered times.** The filtered times are calculated as experimentally measured replication times minus the mean of the approximated replication times. They are compared against a normal distribution with mean = -62.7735 and standard deviation = 113.3735, which is shown as well.

Altogether, our results indicate that DNA replication times are due to a sequence-specific and a sequence-independent part and therefore, they can be split up. Within the sequence-specific part, it is rather the segment length than the segment composition that influences the replication time, which is why the linear model fits almost as good as *model 1*. It seems intuitive that the replication time is longer for larger segments of DNA. Nevertheless, filtering this from the data enabled us to physically locate and map sequence-independent components with a certainty of 95% under the prior normal distribution. Figure 4.4 shows that rate deviations that are caused independently of the underlying sequence, are not scattered randomly across the genome, but are clustered on distinct locations within the genomic landscape of budding yeast. As such, we provide here a map of the regulatory diversity of yeast DNA replication.

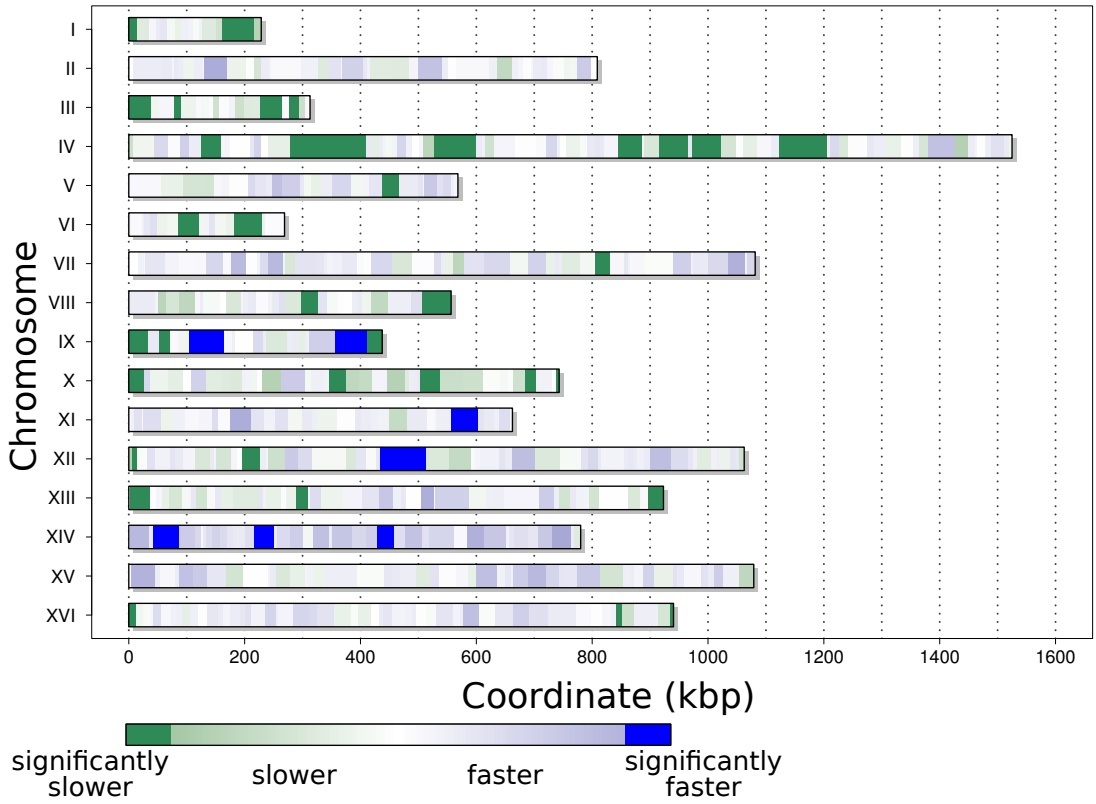


Figure 4.4: **Regions of replication rate deviation for the 16 yeast chromosomes.** Deviations within the filtered times across the genome of budding yeast are shown. Blue shades indicate faster replication in the experiments than predicted by *model 1*, whereas green shades indicate slower replication in the experiments (linear scale, lighter tones indicate smaller deviations). Dark shades indicate a significant deviation from the prior normal distribution. A quantitative view of the deviations (in seconds) for each chromosome can be found in the Appendix C, Figure C17.

4.4 Discussion

In this work we aimed at quantifying effects that influence DNA replication time in budding yeast. We described the movement of the replication machinery along the DNA template as a directed random walk. By using this approach, we decomposed influences on DNA replication time into two major components, a sequence-specific one and a sequence-independent one.

We have shown that the nucleotide composition of a segment does not significantly influence its replication time. Obviously, we cannot rule out completely that there is a nucleotide composition-specific effect on the replication time. It seems intuitive to assume that there are fluctuations, e.g. in the availability of nucleotides in the nucleus. In our analysis, the probabilities p can be viewed as an expression of such fluctuations. They summarize a mixture of factors, incorporating the nucleotide availability among others. However, the contribution of nucleotide composition seems to be too small, at least for the wild type, to be detected by our method using the experimental data taken from Raghuraman et al. (2001). The scenario might be quite different under cellular stress conditions or in nucleotide composition effecting mutants.

We have demonstrated a strong correlation between segment length and replication time. Once again, this seems to be intuitive, since we can assume that the longer a segment is, the more time it will take to be replicated. Nonetheless, we filtered these two results (non-nucleotide-dependency and length correlation) from the replication times. This left us with a distribution of replication times, independent of sequence and length. From the filtered replication times we could directly infer the distribution of replication rates, since all length-specificity is filtered out. This means that, if the replication time is longer than average, the rate would be decelerated and *vice versa*. The distribution of filtered times was then approximated by a normal distribution. We assumed that all deviations from that normal distribution indicated some form of regulation. Applying this logic, we physically located and mapped sequence-independent components with a certainty of 95%. We observed that regions with significant deviations (violating the assumption of normal distribution) do not show uniform spatial distribution but are clustered on distinct locations, which forms a regulatory landscape within the budding yeast genome. Thus, a large part of the elongation time is dictated by some spatial and sequence-independent factors. We therefore, present evidence for another aspect, beyond initiation and origin timing, of the puzzle that is the understanding of regulation of DNA replication in time and space.

What exactly regulates DNA replication in the regions where we observed a significant faster or slower replication (see Fig. 4.4) is not clear. Although, it has been shown that epigenetic factors can influence DNA replication, none of them directly corresponds to the regions we identified (Wintersberger, 2000; Zhang et al., 2000; Ji et al., 2001; Mechali, 2001; Pasero et al., 2002; Antequera, 2004). Nevertheless, an inhomogeneous histone acetylation/methylation pattern could lead to differences in DNA unwinding efficiency, which might cause the observed effect. Histone modification status and remodeling of the chromatin structure could influence the rate at which the replication machinery operates. In fact, particular dense packing of the DNA tertiary structure

could account for deceleration of the replication rate and therefore, modulate origin activity as well (Tabancay and Others, 2006). On the other hand, loosely packed or already unwound DNA, due to e.g. transcription, could facilitate replication (Lucchini and Sogo, 1994; Deshpande and Newlon, 1996; Wellinger et al., 2006). However, it is still under investigation whether these mechanisms of regulation are tightly related to DNA replication or if they are merely the side effects of the regulation of other processes, e.g. transcription. At this point, the reasons for the observed local deviations in the replication times remain unclear, but this might be changed as more and more experimental data becomes available.

There is a number of experiments that could be directly inferred from our results, e.g. transfer a significantly slower or faster replicating segment to another location in the genome and check whether the replication time is conserved, or mutate the sequence of this segment to investigate the potential changes of the elongation time. Considering the tight connection between DNA replication and the other cell cycle events, a link between the replication speed and the accessibility of the origins is likely. In particular, this might be the case for origins that show delayed replication due to the chromatin state of the chromosomes (Tabancay and Others, 2006) or to the Cdk1-Clb5 activity (McCune et al., 2008).

On a different note, in this work we have shown, by using the Akaike Information Criterion (Akaike, 1974), that the replication rate in budding yeast can be best approximated using only a single parameter, as we have proposed in chapter 3 and in Spiesser et al. (2009). Naturally, one could argue that we did only test models that consider sequence-specific attributes and no spatial regulatory events. However, we have shown that spatial regions of interest are not randomly distributed, which is why they can only be described explicitly.

In a further development of the analysis presented, we anticipate to relax some of our modeling assumptions. For example, in budding yeast, polymerases α, δ and ϵ are localized to early firing origin regions during early S phase, suggesting that they function together at multiple replication forks (Hiraga et al., 2005). Their contribution for the apparent speed of the DNA replication process however, has still to be highlighted. In this direction, our study could be suitable for further investigation of their distinctive roles and velocities in the polymerization process. As soon as more experimental data regarding the polymerase kinetics will become available, our model could be extended. In addition, it could be interesting to further investigate stochastic components of DNA replication dynamics in budding yeast. Since S phase dynamics depends both on the replication fork velocity and the initiation frequency of origins (as discussed in chapter 3), an interesting aspect is to combine time-dependent changes in the replication origin activation and a fork density-dependent affinity of the different polymerases for the origins.

5 Different Groups of Metabolic Genes Cluster Around Early and Late Firing Origins

In this chapter, I present a genomic analysis that has been inspired by the idea, outlined in section 3.3.2, that replication origin sequences potentially have evolved to be exclusively sensible for either Cdk1-Clb5 (late firing) or Cdk1-Clb6 (early firing) and that this property might also be mirrored in the sequences (genes) in origin proximity. The chapter is based on:

T. W. Spiesser and E. Klipp. Different Groups of Metabolic Genes Cluster Around Early and Late Firing Origins of Replication in Budding Yeast. *Genome Informatics*, 24(1):179-192, 2010.

5.1 Introduction

DNA replication is a fundamental process that is tightly regulated during the cell cycle (Bell and Dutta, 2002). In budding yeast it starts from multiple origins of replication and proceeds in a timely fashion according to a reproducible temporal program until the entire DNA is replicated exactly once per cell cycle (Alvino et al., 2007; Raghuraman and Brewer, 2010). In this program an origin seems to have an inherent firing probability (see sections 1.2.4 and 3.3.3) at a specific time in S phase that is conserved over the population (Rhind et al., 2010). However, what exactly determines the origin initiation time remains obscure to this day. In the following, we analyze the gene content that clusters around replication origins following the assumption that inherent origin properties that determine staggered initiation times could potentially be mirrored in the close origin proximity due to concomitant sequential evolution. Thus, we collect genes associated with replication origins and perform a gene ontology term enrichment test, as outlined in section 5.2. We find that metabolic genes are significantly over-represented in the regions that are close to the starting points of DNA replication. Furthermore, functional analysis also reveals that catabolic genes cluster around early firing origins, whereas anabolic genes can rather be found in the proximity of late firing origins of replication. In section 5.4 I discuss our findings and speculate that, in budding yeast, gene function around replication origins correlates with their intrinsic probability to initiate DNA replication at a given point in S phase.

5.2 Materials and Methods

Information about origins of replication were taken from the OriDB (Nieduszynski et al., 2007). In this work we have considered all origins that are currently listed (735: confirmed, likely and dubious; February 11th, 2010). Information regarding genomic features of budding yeast were obtained from the *Saccharomyces* Genome Database (SGD) (Cherry et al., 1997) in form of the downloadable SGD_feature.tab and the chromosome_length.tab files (SGD Project). We have identified all verified open reading frames that are located in the vicinity of the origins of replication (target gene set). Herein, vicinity is defined as the region that spans 2 kb up- and downstream of the medial position of the origin (i.e. 4 kb region). A gene is positively identified if the 3' end, the 5' end or the whole gene lies within or stretches over the whole region, as illustrated in Figure 5.1.

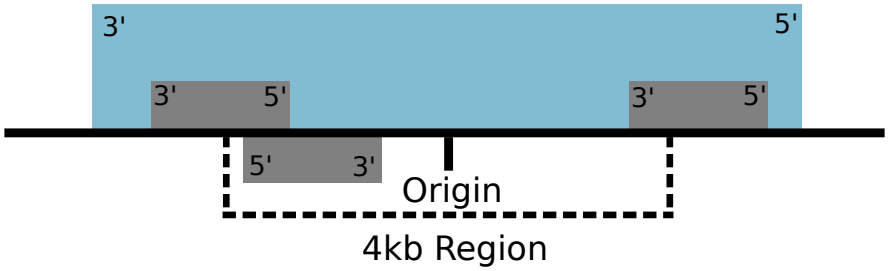


Figure 5.1: **Scheme of the gene-origin association criterion.** The medial position of an origin was chosen to define a 4 kb region around the replication origin on the genome. A gene is associated with this region if the 3' end, the 5' end or the whole gene lies within (shown in gray) or stretches over (blue) the whole region.

Using all verified open reading frames of the budding yeast as a reference set, we performed a functional analysis of the target gene set. The analysis is based on the association of gene ontology (GO) terms (Ashburner et al., 2000) to genes and has been performed using GOstats (Falcon and Gentleman, 2007). GOstats is a package of the *R* statistics environment (R Development Core Team, 2007) and is available from the Bioconductor project (Gentleman et al., 2004). We have tested for over-representation of GO terms in our target gene set by applying a conditional hypergeometric test (Falcon and Gentleman, 2007), with a p-value (p) cutoff of 0.01. The conditioning of the commonly used hypergeometric test corrects for the problem of the hierarchical structure of GO. GO terms usually inherit the annotations from more specific descendants. This often leads to classification of directly related GO terms that have a high degree of gene overlap as being significant at a specific p-value cutoff. The conditioning, implemented by Falcon and Gentleman (2007) solves this problem by removal of all genes that are

annotated at significant children from the gene list of the parent, an approach similar to that proposed by Alexa et al. (2006).

For comparison, we also performed the conditional hypergeometric test on 1000 gene target sets identified on the basis of 735 random locations. A random number generator has been used to randomize the positions of the origins on the chromosomes. However, the origin position change is only allowed within the appropriate chromosome. Thus, the positions of the origins change but the number per chromosome remains the same. The new positions were sampled from a uniform distribution with density:

$$f(x) = \frac{1}{(max - min)} \tag{5.1}$$

for $min \leq x \leq max$. For every chromosome we used $min = 1$ and $max = length\,of\,the\,chromosome$.

Then, we approximated the density distribution function of the p-value distribution, resulting from the 1000 tests using a non-parametric estimator:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5.2}$$

with $x_i, i \in (1, n)$ being the samples of the random variable, a Gaussian kernel K (mean = 0, variance = 1, as defined in equation 1.23) and bandwidth h that is automatically chosen, thus

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}. \tag{5.3}$$

Furthermore, we calculated the ECDF, as defined in equation 1.24, from the distribution of the p-values. Generally, the ECDF can be used to calculate the probability to obtain a certain value (or smaller) under a given distribution.

Finally, we have divided the origins of replication into two different clusters. We classify the origins according to the time at which they initiate DNA replication in a given S phase of the cell cycle. Different studies have identified initiation times for a large number of the origins, nonetheless the information available remains incomplete (Raghuraman et al., 2001; Yabuki and Terashima, 2002). Therefore, we have used the whole genome study by McCune et al. (2008) (for details see sections 1.2.4 and 3.3.2) to classify whether an origin of replication lies within a CDR or non-CDR. This procedure allowed for the separation of the origins into early and late firing origins. We have tested for GO term over-representation of genes associated with both clusters.

5.2.1 Software

All tasks were implemented and analyzed with the programming language *Python* (van Rossum, 1995) and the *R* statistics environment (R Development Core Team, 2007). *Rpy*, a high level *Python* module for managing the lookup of *R* objects, has been used for the internal communication between *Python* and *R*.

5.3 Results

Table 5.1 shows significant hits of the GO term enrichment analysis for genes that are located in the vicinity of origins of replication. In this work, we used all verified open reading frames of *S. cerevisiae* (4844) as a reference set (gene universe) for the conditional hypergeometric test and analyzed 1388 genes located in the immediate origin local area. 21 terms have been identified using a p-value cutoff of 0.01 and an origin vicinity margin of 4 kb.

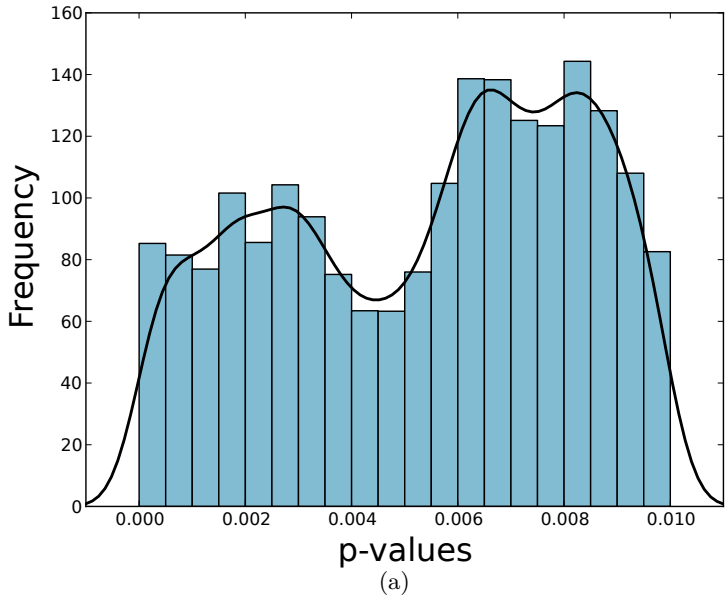
We have further subcategorized the enriched GO terms. The first and largest of the subgroups represents metabolic processes. 10 out of the 12 most significant hits fall into that category, e.g. alcohol catabolic process ($p \sim 2.15422 \cdot 10^{-6}$) or thiamin biosynthetic process ($p \sim 0.00077$). Directly related to metabolic processes is the category transport of metabolites. This group represents functional enriched genes of carbohydrate, hexose and glycerol transport annotations. The third group contains genes of cell cycle processes and development, e.g. synaptonemal complex assembly ($p \sim 0.00879$) and the fourth group RNA processing genes. Response to toxin ($p \sim 4.61234 \cdot 10^{-6}$) could not be assorted to any of the categories.

For comparison we have performed 1000 GO term enrichment analysis tests for genes located near origins with random (uniformly distributed) positions. Exemplarily, we present the result of one random test, where we identified 11 terms with a p-values below 0.01 (Tab. 5.2). Five of the terms concern various types of regulation, e.g. positive regulation of organelle organization ($p \sim 0.00106$) or positive regulation of glucose metabolic process ($p \sim 0.00829$). Three terms assort to metabolic processes and the last three terms concern response to copper ion, flocculation and mitochondrial genome maintenance. The p-values are generally higher than the ones determined using the original origin positions. The density distribution of the p-values (1000 tests) and the associated ECDF are displayed in Figure 5.2. The density distribution shows an almost bimodal shape with peaks near 0.003 and 0.008, whereas the ECDF increases nearly linearly. The ECDF obtained using the p-values from the original origin positions is also shown (Fig. 5.2 (b)). It increases in the first half in a saturated curve-like manner and then converges into linear growth in the latter half.

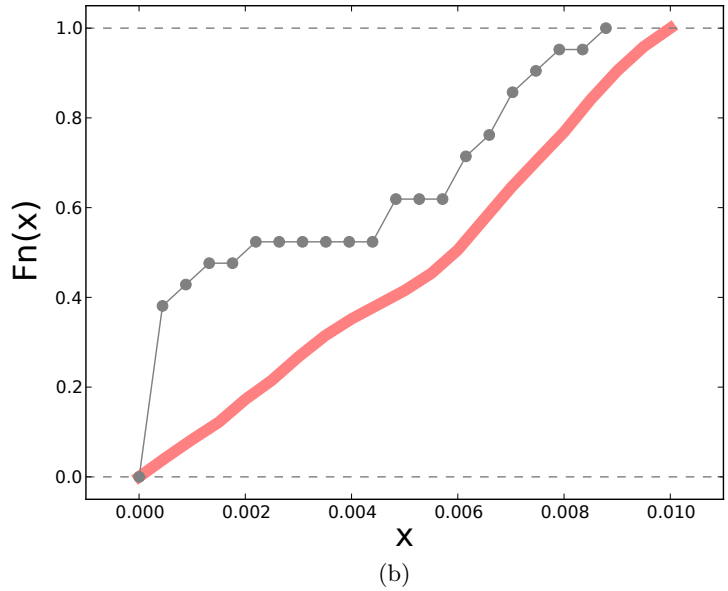
We divided the origins into clusters of early and late replication to study whether different groups of genes are replicated at distinguishable times in S phase. Tables 5.3 and 5.4 show the results for the GO term enrichment analysis for genes in early (non-CDR) and late (CDR) replicating regions, respectively. We found 16 enriched GO terms for 558 genes that are associated with early firing origins. Remarkably, more than half of them (9 out of 16) are related to catabolic processes, e.g. organic acid catabolic process ($p \sim 0.00049$) or aromatic compound catabolic process ($p \sim 0.0079$). Two are associated with metabolic or biosynthetic processes, two with RNA processing, two with DNA packing and one with organelle inheritance ($p \sim 0.0059$).

Term	Count target set	Count reference set	p-value	GOBPID
alcohol catabolic process	40	73	$2.15422 \cdot 10^{-6}$	GO:0046164
hexose catabolic process	34	60	$4.61234 \cdot 10^{-6}$	GO:0019320
response to toxin	24	38	$9.44885 \cdot 10^{-6}$	GO:0009636
monocarboxylic acid metabolic process	68	151	$1.00416 \cdot 10^{-5}$	GO:0032787
monosaccharide metabolic process	63	142	$3.75132 \cdot 10^{-5}$	GO:0005996
carbohydrate catabolic process	42	89	0.00014	GO:0016052
gluconeogenesis	19	31	0.00015	GO:0006094
glycolysis	20	34	0.00022	GO:0006096
thiamin biosynthetic process	13	20	0.00077	GO:0009228
thiamin and derivative metabolic process	14	23	0.00125	GO:0042723
carbohydrate transport	19	36	0.00187	GO:0008643
carboxylic acid catabolic process	26	57	0.00456	GO:0046395
endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA from (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	15	28	0.00473	GO:0000472
endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	14	26	0.00594	GO:0000480
hexose transport	14	26	0.00594	GO:0008645
cellular developmental process	132	384	0.00636	GO:0048869
glycerol transport	4	4	0.00672	GO:0015793
cell wall organization	89	248	0.00672	GO:0007047
ncRNA 5'-end processing	15	29	0.00725	GO:0034471
reproduction of a single-celled organism	79	218	0.00787	GO:0032505
synaptonemal complex assembly	6	8	0.00879	GO:0007130

Table 5.1: **GO term enrichment analysis results for 1388 genes associated with origins of replication.** GO terms, count of genes in target set, count of genes in reference set, p-values (rounded up) and GOBPIDs are shown for significantly enriched terms.



(a)



(b)

Figure 5.2: **Distribution of p-values obtained from 1000 enrichment tests using random locations.** (a) Frequencies are shown as histogram, approximated density distribution is shown as black line. (b) Empirical cumulative distribution function is shown for random location p-values (solid orange line) and for the p-values obtained from testing the original positions of replication origins (shown in gray).

Term	Count target set	Count reference set	p-value	GOBPID
positive regulation of organelle organization	9	12	0.00106	GO:0010638
regulation of gene-specific transcription	17	31	0.00168	GO:0032583
cellular amino acid derivative biosynthetic process	14	25	0.00332	GO:0042398
positive regulation of specific transcription from RNA polymerase II promoter	10	16	0.00449	GO:0010552
hydrogen peroxide metabolic process	4	4	0.00643	GO:0042743
response to copper ion	4	4	0.00643	GO:0046688
flocculation	7	10	0.00748	GO:0000128
mitochondrial genome maintenance	16	32	0.00759	GO:0000002
positive regulation of glucose metabolic process	6	8	0.00829	GO:0010907
positive regulation of carbohydrate metabolic process	6	8	0.00829	GO:0045913
amine catabolic process	21	46	0.00894	GO:0009310

Table 5.2: **Exemplary GO term enrichment analysis results for genes associated with 735 random locations on the genome.** GO terms, count of genes in target set, count of genes in reference set, p-values (rounded up) and GOBPIDs are shown for significantly enriched terms.

Concerning the 773 genes that are localized close to late firing origins, we found 30 enriched GO terms (Tab. 5.4). 14 terms are related to various kinds of metabolic processes, e.g. vitamin metabolic process ($p \sim 7 \cdot 10^{-5}$) or gluconeogenesis ($p \sim 0.00623$), 9 terms represent genes that we classified as cell cycle and development related, as e.g. developmental process ($p \sim 0.00076$) or meiosis I ($p \sim 0.00908$), 6 terms concern genes of compartmentalization, e.g. cell wall organization ($p \sim 0.00192$) or spore wall biogenesis ($p \sim 0.00235$) and one term represents genes that are involved in the transport of glycerol ($p \sim 0.00064$).

5.4 Discussion

In this chapter, I present the results of a functional analysis of genes that we found to be positioned close to origins of replication. A conditional hypergeometric test was used to cluster functionally related genes according to their GO terms and to determine significant over-representation. We found that genes related to metabolic processes were most prominently over-represented amongst the genes that were tested (10 out of the 12 best hits, see Tab. 5.1). We calculated p-values that could be expected by chance, using the results of 1000 tests with randomized positions and the probabilities of the p-values obtained from the original test. The probabilities to obtain the p-values of the first 8 hits are around 1%. This means that the odds to obtain such an association by chance lie around 1%.

In addition, the gene target set has been split to test whether different groups of genes cluster around early and late firing replication origins. Genome-wide data concerning the dependency of replication times on Clb5 was used to classify the genes to either lie in early or late replicating domains (McCune et al., 2008). Figure 5.3 shows all genes that have been identified to be in the vicinity of origins, using a vicinity margin of 4 kb on a genome scale, where replication origins, CDRs, non-CDRs and inconclusive regions are indicated as well. Since origins, as well as genes, occupy a certain terrain on the genome, it seems apparent that a gene could generally be classified to belong to more than one origin region. Theoretically, the two origins could lie on the border of a CDR and a non-CDR, so the gene in question could, in that particular case, not unambiguously be assigned to be located in an early or late firing domain. In order to test for this special case, we investigated how many genes allocate to more than one origin. We found this to be true for 107 genes. Consequently, we further tested how many of them could potentially fall into both (CDR and non-CDR) regions and detected this to be the case for only three genes (YLR081W, YMR246W and YER136W). Hence, the three ambiguous genes have not been considered in the analysis. Furthermore, it has to be mentioned that one gene (YAR020C) lies within a region that was termed 'Inconclusive' by McCune et al. (2008) and is therefore, neither classified as early nor late. Thus, we did not consider YAR020C in the analysis of the early and late replicating domains either. Additionally, the analysis of McCune et al. (2008) does not give any information regarding the first and last 12 kb of every chromosome, which gives a total of 57 genes (including the ones mentioned above) that have not been considered in the analysis of early and late replicating domains.

We found 16 GO terms for functionally enriched genes close to early and 30 terms for genes close to late replicating origins. Genes related to metabolic processes also dominate the GO terms in both domains when separated. However, it seems that metabolic genes that cluster around early origins mostly concern catabolic reactions (9 out of 11). Intrigued by this, we investigated the metabolic genes around late origins in more detail as well. 14 terms relate to metabolic processes, where 7 of them cannot be distinguished on first sight (e.g. vitamin metabolic process), 5 concern anabolic reactions (e.g. thiamin biosynthetic process) and two of them catabolic ones (e.g. hexose catabolic process). Therefore, we investigated the structure of the GO tree around the 7 indistinguishable

metabolic terms and the genes in the gene target set that relate to them. We found that the vitamin metabolic process (74 genes) has the following four children: regulation (2 genes), water-soluble vitamin metabolic process (70 genes), biosynthetic (64 genes) and catabolic (1 gene) in budding yeast. A closer look into our gene set told us that the catabolic gene is not part of our gene set. Furthermore, since we applied the hypergeometric test with the conditional correction and water-soluble vitamin biosynthetic process (a child of water-soluble vitamin metabolic process) is a significant term and therefore, taken out of the set when testing the vitamin metabolic process, it follows that the majority of genes to be tested must be out of the 64 annotated biosynthetic genes. Thus, we conclude that the vast majority of vitamin metabolic process genes actually concerns anabolic reactions in the target set, since no catalytic ones could be found and 64 out of 74 are anabolic. The same procedure has been applied for the other 6 indistinguishable metabolic terms. It became apparent that also for thiamin (vitamin B1) and derivative metabolic process, pyridine nucleotide metabolic process, NADP metabolic process and alkaloid metabolic process no catalytic genes were in the gene target set. Regarding monocarboxylic acid metabolic process and coenzyme metabolic process we could not fully determine the single contributions of our gene target set due to complexity of the gene composition concerning those GO terms. A more sophisticated method needs to be developed in the future to investigate those nondistinctive terms. Nonetheless, it seems that, in budding yeast, catabolic genes cluster around early and anabolic genes around late origins of replication.

We speculate that this phenomenon might be the results of an evolutionary optimization designed to cope with the increasing costs during cell division. The early replication of catabolic genes results in early duplicates of those genes, which increases their transcriptional capacity and thus, potentially their mRNA levels as well (sketched in Fig. 5.4). Consequently, cells that double their catabolic genes in early S phase can benefit much longer from a potentially heightened catabolic capacity. This could potentially lead to a shift of the metabolic rate and with it to a shift of the growth rate. In chapter 2, we have already presented experimental evidences from various sources showing that the growth rate changes in the course of the cell division cycle, particularly at the beginning/mid S phase (Aldea et al., 2007; Cookson et al., 2009; Goranov et al., 2009). While the reason for this rate shift remains elusive, it has been speculated that it could be due to a, through DNA replication induced, natural gene-dosage effect (Mitchison, 2003). Indeed, such an effect has already been described more than fifty years ago and was then termed rate changing point (Mitchison, 1958). The results presented in this chapter specify the gene-dosage hypothesis in regard to its timely occurrence and fine-tuning. Not only does it seem that the gene-dosage of the entire genome result in increased growth, but also its time-resolved process seems to be fine-tuned to optimize growth. In the particular case of a natural gene-dosage effect, the genomic position can function as a modifier of gene expression.

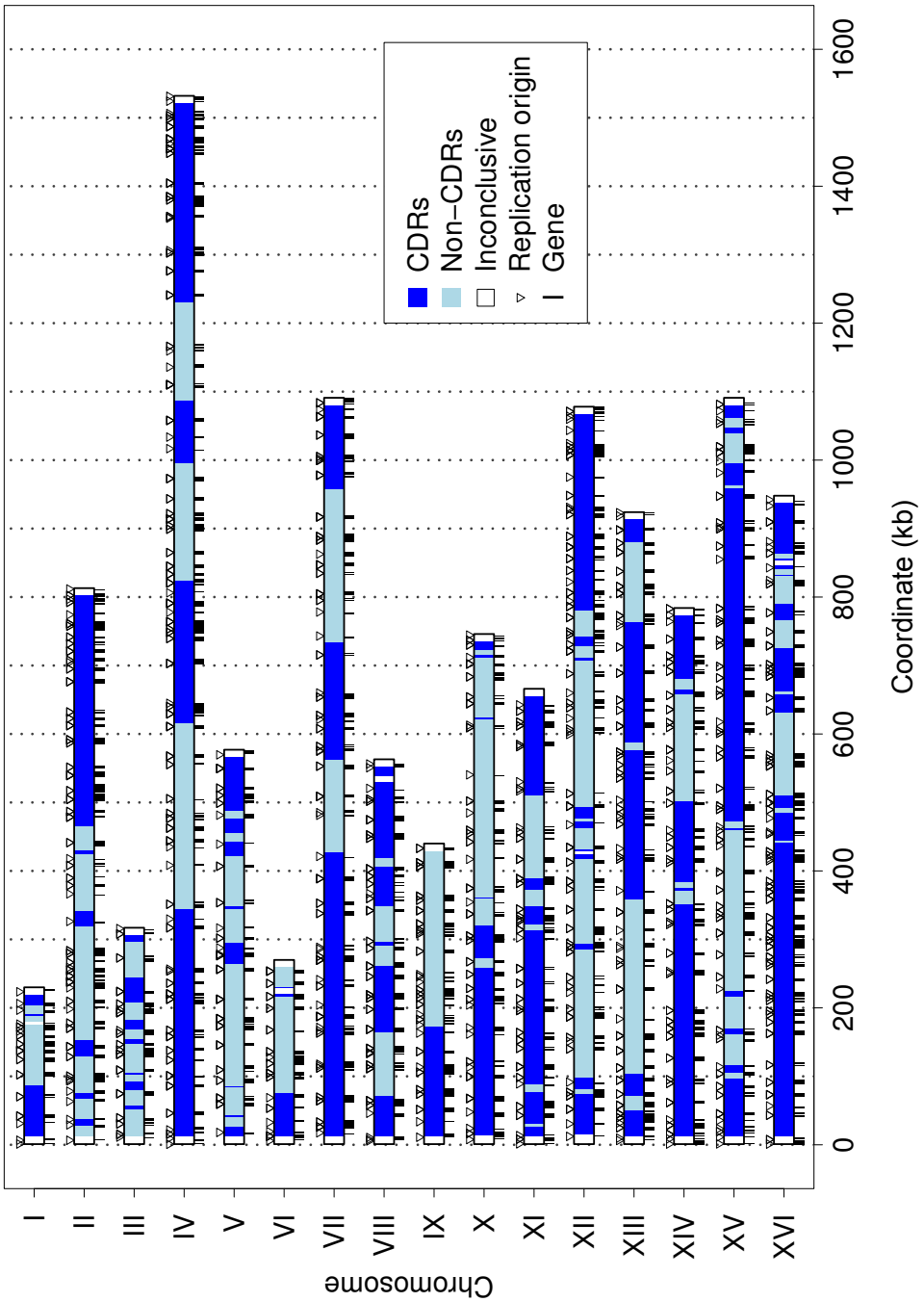


Figure 5.3: **Chromosomal location of replication origins (triangles) and associated genes in origin vicinity (black lines).** Furthermore, CDRs (blue), non-CDRs (light blue) and Inconclusive regions (white) are shown, as identified by McCune et al. (2008).

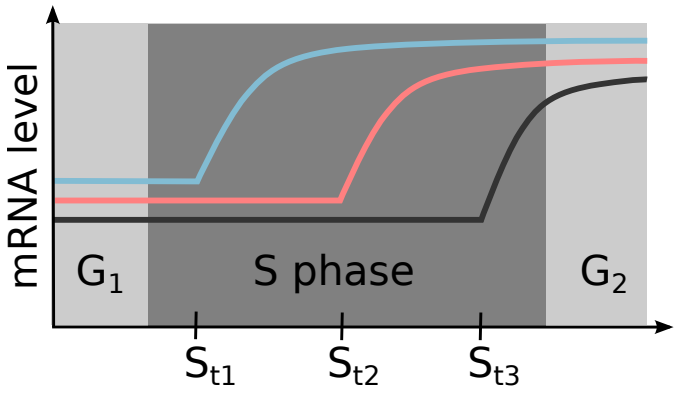


Figure 5.4: **Schematic mRNA levels during S phase.** Given a gene is transcribed with rate k_1 and the resulting mRNA is degraded with rate k_2, it holds that the mRNA level is constant after a while. If a gene is replicated early in S phase (S_{t1}) the gene itself and its copy can be transcribed until the cell finally divides. This doubles rate k_1 which results in increased mRNA levels. Accordingly, mRNA levels are affected later and thus, shorter, for genes that are replicated late in S phase (S_{t3}). Thus, location of a gene can influence its expression during S phase.

In conclusion, we found that especially metabolic genes are localized close to replication origins. Probabilities for such highly significant over-representations have been calculated using the probability distribution that could be obtained by random location tests. Under the assumption that certain origin properties, such as probabilities for early or late initiation, could potentially be mirrored in the origin environment, we separately tested genes in early and late firing domains according to functional over-representation. Indeed, apart from chromatin status and correspondingly transcriptional activity, two factors that are most closely connected with origin activation time *per se*, also the gene function around origins seems to reflect some basic property of DNA replication. That is to say that metabolic genes near early firing origins mostly concern catabolic reactions and the majority of the metabolic genes near late firing origins are responsible for anabolic processes. It is tempting to speculate that origins and gene sequences in their close proximity might have evolved through e.g. duplication events, optimizing energy allocation and conserving inherent properties of a particular genomic region along the way.

Term	Count target set	Count reference set	*p*-value	GOBPID
organic acid catabolic process	16	57	0.00049	GO:0016054
nucleosome assembly	10	29	0.00099	GO:0006334
nitrogen compound catabolic process	16	61	0.0011	GO:0044270
threonine catabolic process	3	3	0.00152	GO:0006567
maturation of 5.8S rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	17	69	0.00165	GO:0000466
DNA packaging	17	70	0.00195	GO:0006323
pyruvate metabolic process	12	42	0.00206	GO:0006090
cellular amino acid catabolic process	12	42	0.00206	GO:0009063
monosaccharide catabolic process	16	68	0.00369	GO:0046365
allantoin catabolic process	4	7	0.00458	GO:0000256
glucose catabolic process to ethanol	4	7	0.00458	GO:0019655
organelle inheritance	15	65	0.0059	GO:0048308
endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	8	26	0.00697	GO:0000480
aromatic compound catabolic process	5	12	0.0079	GO:0019439
catabolic process	121	866	0.00835	GO:0009056
alcohol biosynthetic process	17	81	0.00952	GO:0046165

Table 5.3: **GO term enrichment analysis results for 558 genes associated with early firing origins** of replication (positioned in non-CDRs). GO terms, count of genes in target set, count of genes in reference set, *p*-values (rounded up) and GOBPIDs are shown for significantly enriched terms.

Term	Count target set	Count reference set	p-value	GOBPID
vitamin metabolic process	32	102	$7{\cdot}10^{-5}$	GO:0006766
thiamin and derivative metabolic process	11	23	0.00034	GO:0042723
thiamin biosynthetic process	10	20	0.0004	GO:0009228
glycerol transport	4	4	0.00064	GO:0015793
developmental process	96	447	0.00076	GO:0032502
hexose catabolic process	19	60	0.00176	GO:0019320
cell wall organization	57	248	0.00192	GO:0007047
cell differentiation	55	239	0.00223	GO:0030154
ascospore wall assembly	15	44	0.00235	GO:0030476
spore wall biogenesis	15	44	0.00235	GO:0070590
cell wall assembly	15	44	0.00235	GO:0070726
fungal-type cell wall biogenesis	17	54	0.00324	GO:0009272
regulation of cell division	6	11	0.00363	GO:0051302
reproduction single cell organism	50	218	0.00374	GO:0032505
alcohol catabolic process	21	73	0.00386	GO:0046164
medium-chain fatty acid biosynthetic process	3	3	0.00405	GO:0051792
water-soluble vitamin biosynthetic process	18	60	0.00445	GO:0042364
reproductive process	42	178	0.00446	GO:0022414
pentose-phosphate shunt	7	15	0.00508	GO:0006098
pyridine nucleotide metabolic process	16	52	0.00545	GO:0019362
gluconeogenesis	11	31	0.00623	GO:0006094
M phase of meiotic cell cycle	44	192	0.00644	GO:0051327
monocarboxylic acid metabolic process	36	151	0.00686	GO:0032787
coenzyme metabolic process	35	146	0.00697	GO:0006732
ascospore formation	28	111	0.00733	GO:0030437
sexual sporulation	28	111	0.00733	GO:0034293
premeiotic DNA synthesis	4	6	0.00737	GO:0006279
NADP metabolic process	9	24	0.0087	GO:0006739
alkaloid metabolic process	15	50	0.00905	GO:0009820
meiosis I	22	83	0.00908	GO:0007127

Table 5.4: **GO term enrichment analysis results for 773 genes associated with late firing origins** of replication (positioned in CDRs). GO terms, count of genes in target set, count of genes in reference set, p-values (rounded up) and GOBPIDs are shown for significantly enriched terms.

6 Discussion and Concluding Remarks

The objective of this thesis was to explore two aspects that are crucial for the proper progression of the growth and division cycle. Those aspects are the mechanisms (i) employed to maintain size homeostasis during G_1 and (ii) of DNA replication during S phase of the cell cycle. Inaccurate size control as well as incomplete DNA replication lead to checkpoint activation which induce cell cycle arrest. Defects in both systems and their control points have been related to cancer (Hanahan and Weinberg, 2011). I have presented detailed mathematical formulations for both processes developed for the premier model organism budding yeast. Herein, the systemic approach was grouped into four main research projects, namely the study of (1) size regulation and homeostasis of yeast cell populations, (2) the spatiotemporal organization of DNA replication, (3) differences in the DNA replication machinery motion (elongation) and (4) the functional association of genes and replication origins.

For (1) we have created a single cell ODE core model that is complemented with a stochastic component. We deduced population behavior from the single cell model through multiscale simulations using an environment that we especially developed for this purpose. For (2) we implemented an algorithm that simulates the DNA replication process. We used this algorithmic model to test the impact of different replication origin activation patterns. (3) was assessed with a fine-grained stochastic model for the replication machinery motion along the DNA template strand and for (4) we used hypergeometric gene ontology association tests.

The main findings of the projects presented in this thesis are (1) that size regulation is an intrinsic property of yeast cell populations and that no signaling or size sensing mechanism is required for it, (2) that DNA replication is robust against perturbations, especially in small chromosomes with high origin density, (3) that there are distinct locations in the genome where the elongation process is strongly biased and (4) that catabolic genes are over-represented near early origins and anabolic genes near late origins.

The cyclic systems biology approach - a good way to do research?

The systemic research approach presented here was conducted in an iterative model construction/refinement and model - data comparison cycle, as representatively illustrated by the idealized systems biology workflow shown in the introduction (see section 1.3.2). Throughout the iteration process, hypotheses were tested and the knowledge about the specific systems was refined. However, in comparison to the idealized process cycle, real life research is different. Scientific progress is rarely strictly cyclic nor is it straightforward, but it is rather achieved through moving back and forth in different directions

through the cloudy mist that is the unknown (Alon, 2009). Consistently, we still have, in the course of model construction, iteratively compared simulations with experimental data of various sources with one another and refined/redesigned the models accordingly. Nonetheless, only the final models are presented here, whereas earlier/different versions of the models are generally not shown, except for when different model versions are explicitly compared to one another, as in chapter 4.

The iteration is a crucial part since it helps understanding the biological system further and thus, helps gaining knowledge. Therefore, also in future theoretical studies we hold on to the systems level research strategy. However, it is noteworthy that the biological value of the theoretical work extends, when model predictions help guiding experiments or provide a benchmark against which experimental data can be tested. In this manner, the model as well as the generated data can eventually be evaluated. So far, we have exclusively used published experimental data for model - data comparison (for example Cookson et al. (2009) in chapter 2). While this approach is valid during model construction and validation as long as appropriate data is available, a key feature of a model remains its predictability and therefore, to close the systems biology cycle, we envision for the future to test some of the model predictions.

Future prospects and methodology

Several testable hypotheses have been generated, among them (a) the noise reduction relative to average size at high growth rates described in chapter 2, (b) segment specific elongation rates assessed in chapter 4 or (c) that a rate changing point, already described by Mitchison (1958, 2003), could be due to differential replication time of anabolic and catabolic genes, as proposed in chapter 5. We also proposed different experiments that are directly inferred from the results to test our hypotheses. For (a) to measure the cell sizes in populations grown on different media, validate/falsify (b) by mutating the sequence of particularly fast/slow replicating segments to assess if there are changes of the elongation time and finally, for (c) to investigate expression levels of an especially early replicating catabolic/late replicating anabolic gene product as a showcase.

From a theoretical point of view there are also a number of results that deserve follow-up studies. For instance, we have proposed in chapter 2 that the G_1 network, that was generally thought to be implicated in setting the critical size threshold at START (Barberis et al., 2007) is dispensable for size regulation in yeast populations. Hence, it would be intriguing to extend our model with the G_1 network components. Such an extended version could be used to study the effect of the network on size regulation, noise reduction or a size related impact of cellular stress responses. Others would be to implement a more sophisticated stochastic representation for the transcriptional process. It has been shown that transcription in single cells occurs generally in large bursts (Elowitz et al., 2002; Kaufmann and van Oudenaarden, 2007), as we have implemented it. However, studying different modes of burst generation and the subsequent noise propagation to downstream targets and thus, on size regulation, might be an interesting starting point for future modeling. Moreover, stochastic simulation based on 3D diffusion of particles could be used for low amount quantities, e.g. mRNAs (Gillespie, 2007). In this manner,

the impact of noisy transcription and noise propagation on cell size distributions could be tested directly (Bruggeman et al., 2009). In the cases of low molecular abundance it is very likely that fluctuations are influential and stochastic descriptions could provide new insights into how the cell size distributions and homeostasis is shaped by it. Nonetheless, since the size regulation system is highly dynamic, also in future implementations dynamic modeling methods probably remain the methodology of choice (see section 1.3.2).

For simulating the process of DNA replication usually algorithms are used (Goldar et al., 2008; Brümmer et al., 2010). Although, the mode of origin activation can be represented differently - deterministic as in Spiesser et al. (2009) or stochastic as presented by Yang et al. (2010) - once activated the elongation is assumed to proceed continuously. We have tested this assumption with a detailed stochastic interpretation of the elongation and could show that the global elongation behavior over a population of cells is best approximated with a single value (Spiesser et al., 2010). Although, in essence correct, on the single cell level the stochastic description remains indispensable. The reason for this is, and here replication and size regulation overlap, that transcription as well as translation occur in bursts that are shaped by elongation dynamics (Dobrzynski and Bruggeman, 2009). Therefore, it seems likely that also the replication process is essentially governed by elongation dynamics in single cells. Thus, also in future implementations of DNA replication, we maintain the current methodological point of view and use either a global average for population dynamics or detailed stochastic models for the study of elongation dynamics in single cells. It will be interesting to further study elongation as a stochastic process. For it, one could test if theoretical formulations of DNA replication with elongation occurring in bursts are compatible with empirical data as well. A potential target for proteins that cause bursts through collision and pausing, such as ribosomes in translation, are Mcm2-7 helicase molecules. They have been shown to assemble in excess at the initiation sites and their motion could well be influenced by collisions and pausing as well (Lei et al., 1996; Hyrien et al., 2003). Our model (3) seems most suitable for a minimal extension in order to conduct this particular experiment.

Choosing different perspectives

On a different note, I would like to draw the attention to a feature of systems modeling which is nicely illustrated with the help of this thesis. It is the fact that the use of different scales serves to explore different levels of a problem. This is required for an integrative view and to increase the understanding of a system as a whole. Different levels can be e.g. single pathways, cells, populations, organs, tissues, an organism or even entire ecological webs and every level gives rise to another perspective, shedding light on a problem from another angle (Ideker and Lauffenburger, 2003). Only when different perspectives are combined we can learn to understand systems in all their complexity. That is also to say that choosing only one angle from which to study a system, might suffice to obtain a general overview of the system, but it might not be enough for a complete understanding of all mechanistic properties and interactions in complex systems. For that, information from different perspectives and angles need to

be integrated, dynamics of different levels need to be explored and data from multiple scales, resolutions and modalities needs to be integrated (Kitano, 2002). This task is non-trivial. Still, it is easily demonstrated when considering the following example from project (1) shown in chapter 2.

We model growth and division for a single cell and use ODEs for the description and to follow the time-dependent evolution of the single components in the cell. In this case the cell is the complex system and the perspective is the focus on the dynamics of the cellular components. We cannot intuitively predict the behavior of the cell and only learn that the cells do not show size regulation on the single cell level after simulations of the model. If one was to stop at this junction, one would probably conclude that we must have overlooked some form of size sensing mechanism that measures a critical cell size for them to commit to cell division, a hypotheses that has been general consensus for many years (Alberghina et al., 2004; Dez and Tollervey, 2004; Cook and Tyers, 2007). However, when zooming out and changing the perspective, more insights can be gained and also the reasoning changes. In our case, we looked at the population behavior level, which means that now the population is the complex system and the single cells are the components. One could argue that our components do not interact and therefore, the population is not a complex system. However, cells give birth to new cells and the initial conditions of the new daughters are shaped by the states of the mothers. Thus, the dynamics of the single components are interlinked. The evolution of the population is driven by the daughters which are shaped by the dynamics of the mothers, which is why, again, the nature of this complex system cannot be predicted intuitively. Summing up, we found that size regulation on the single cell level is not needed for population size regulation (section 2.3.3) rejecting the hypothesis that for population size homeostasis, cells need a size sensing mechanism.

Moreover, as introduced in section 1.3.2, studying different aspects of the same system (projects (2-4)) nicely serves as an illustration that the choice of the modeling formalism defines the granularity of the systemic property that can be studied and that *vice versa* the problem dictates the formalism that is to be used for its representation.

The field of size regulation - how do we contribute?

Ramanathan and Schreiber (2007) state about the dynamical system underlying size regulation that: “Cell growth is, in general, regulated by a linkage between growth rate, cell size and cell division”. Some properties of this dynamic system are well known, such as that cells grown on richer media grow faster and become large, whereas cells grown on a poor medium grow slower and remain in comparison quite small (Tyson et al., 1979). Here, we have assembled as many of these properties that we could find in the literature to paint a concise qualitative picture of the system that determines cell size and with it, to provide a benchmark against which growth models can be validated (Tab. 2.4). For example, it is also well known that yeast cells divide asymmetrically and as such that there are different G_1 phase lengths in mothers and daughters to compensate for the resulting size difference (Hartwell and Unger, 1977; Brewer et al., 1984). Furthermore, single cell sizes are highly divergent in a cell population, due to a high degree of noise

in the system, but also due to the fact that single cells get larger with age (Egilmez et al., 1990; Di Talia et al., 2007). Nonetheless, there is population size homeostasis at a size level that is characteristic for the distinct environmental conditions (Johnston et al., 1977). Moreover, cells grow at different rates in different cell cycle phases as well (Goranov et al., 2009) and on top of this, most G_1 network mutants are viable and still exert size regulation (Enserink and Kolodner, 2010). Given this complexity, neither pinpointing the systemic linkage nor predicting systems dynamics and properties is a trivial task, or as Cook and Tyers (2007) formulate it

> "The proportional control of size, whether for a single cell or an entire organism, is a paradigmatic systems-level problem in biology."

As such, there are long standing questions concerning size regulation that await answering. For instance, is there an inherent size sensing mechanism that the cell employs to set the critical cell size required to commit to division? If yes, what is it and how does it work? Another intriguing question would arise in case cell size is deregulated, resulting in cells that are extremely big or small. What would be the primary goal for the cell in this particular case, regain a reasonable size itself or make sure to produce reasonably sized offspring? In my opinion, this is a just question, since it is the offspring that eventually shapes the face of an exponentially growing culture and not the mother cells themselves (Appendix A, Fig. A2).

With the work, outlined in this thesis, we contribute to the long standing discussion about understanding if and how a cell measures its size and how it knows when to divide at the given specific growth rate. We study the systemic linkage of growth rate, cell size and cell division with a computer modeling approach, in which single cell and population behavior is continuously monitored. We present, for the first time, theoretical evidence substantiated and validated with empirical data of multiple scales (single cell and population data) that cells do not need a size sensing mechanism to exhibit population size homeostasis. Despite the lack of any sensing, regulatory feedback or signaling mechanisms the model is stable over a wide range of growth conditions, exhibiting the above mentioned rate specific size levels (section 2.3.4). Size regulation emerges from the dynamic system as a result of the linkage between metabolic capacity, the cell size and cell division. In contrast to the view, that growth parameters are regulated by the cell cycle, it rather seems to be the other way around. The balancing act between positive regulator: metabolic capacity, and negative regulator: current cell size, simultaneously determines the appropriate cellular growth rate and gates the START transition. As such cell cycle progression and cell size are common output, not input parameters that are regulated in parallel. This explanation might also help understanding the wide tolerance for variations in cell size (Cook and Tyers, 2007).

In response to the second question, we also study the model's robustness against perturbations in initial conditions (section 2.3.5), also including the case that cells are extremely big or small. Here, we find that cells behave in an unexpected, almost altruistic fashion. They balance their growth parameters to equip their daughters with initial conditions much more adapted to the environmental conditions than their own. They do not pollute the daughters with their own deregulated condition, e.g. a big portion of their

own size, in case they are extremely big, just to regain an appropriate size themselves, but divide only when size and metabolic capacity are balanced to produce reasonably sized offspring. The mothers remain in their extreme condition. This remarkable display allows the culture to reach its growth rate specific cell size distribution within very few generations (section 2.3.5, Fig. 2.14).

In summary, understanding size regulation and size regulatory systems dynamics is, as mentioned, not trivial. Here, we provide a minimalistic model that can be used to study size regulation and serves as a starting point for further studies of cell cycle and size related mechanisms on the single cell and on the population level.

The field of DNA replication - how do we contribute?

With one of the major results of the size regulation project, i.e. cell cycle transition and size regulation is regulated in parallel gated by metabolic capacity, we could substantiate previous suggestions in this direction (Bernstein et al., 2007). Thus, once cells have gained appropriate metabolic power during G_1, they also attained a reasonable size and pass START to enter S phase. In S phase, the primary event is DNA replication. Among others, it is the major process required for duplication, since the prerequisite for reproduction and transmitting genomic information to the offspring is exact and efficient replication of the genome. It is a highly controlled cellular process, which makes up a large part of the cell cycle. Severe malfunctions within DNA replication are usually lethal. As such, DNA replication is subject to a complex regulation in all eukaryotic organisms, which makes the identification of the underlying mechanism a non-trivial task.

General replication discussions often concern the understanding of the onset of genomic duplication and its timely organization. Origins initiate replication throughout S phase but there is an ongoing debate about the general mode of initiation. Two opposing schools emerged over the years, arguing for two different points of view. One favors the notion that origin initiation is essentially deterministic (Raghuraman and Brewer, 2010). The other argues that the nature of origin firing is essentially stochastic (Rhind et al., 2010). However, both schools agree on the fact that there are early and late initiating origins, i.e. that there is some form of replication program. Furthermore, as aptly formulated by Rhind et al. (2010) and mediating between both schools

> "[...] in a trivial sense, all models are stochastic. The real question concerns the degree of stochasticity and whether the stochasticity itself plays an important role in replication control. Thus, one might, more loosely, call a model deterministic if the variation in origin firing times is much less than the duration of S phase and stochastic if they are a substantial fraction."

Nonetheless, the issue is still under debate and thus, it is interesting to study and analyze the intrinsic robustness and dynamics that the system displays from both perspectives. In this way, one might be able to narrow down which one of either concepts leads to the displayed dynamics and inherits specific deterministic or stochastic properties to the system. Moreover, investigating evolutionary aspects could prove useful to provide hints

and clues as to why and how the deterministic/stochastic concept has evolved in that particular manner.

To contribute in solving these issues, we studied DNA replication on three different levels. First, we screened replication dynamics and robustness from a deterministic point of view and with it provided the first mathematical description of DNA replication in budding yeast at that time (Spiesser et al., 2009). However, already there, we started to explore stochastic influences on its dynamics (project (2)), which was later taken on by others (Yang et al., 2010; Koutroumpas and Lygeros, 2011). A lasting point of debate in our study remained the approximation of the rate for the elongation process with a constant value. Thus, in a follow-up, we conducted a study for the elongation process and its stochasticity in detail on a genome scale (project (3)). We could show that elongation is remarkably uniform for most of the genome and thus, a single value suffices to best approximate the elongation process. This finding was also later supported by an experiments and modeling study from Sekedat et al. (2010). Finally, to complement our integrative research, we decided to approach the intriguing question of the evolution of the course of events into the program it is today, from a bioinformaticians perspective. We reasoned that genes close to origins would replicate first and speculated about what kind of genes that would be. This is for a simple reason, i.e. that the gene copy numbers can alter gene expression, also known as a gene dosage effect, which can play an important role, e.g. in cell cycle progression as shown by Di Talia et al. (2007). If the DNA replication induced gene dosage effect is important, that is to say that if it plays a role at all, as has been suggested (Mitchison, 1958; Aldea et al., 2007), then there might also be a difference between genes in early or late replicating domains. From a reverse engineering point of view answering this question is intriguing. Could one predict if an origin initiates replication early or late on the basis of the genes in its vicinity? To contribute to answering this, we analyzed the gene content around origins of replication to investigate the sequential evolution of origins and genes in their vicinity (project (4)). Here, we could show that catabolic genes are over-represented near early origins and anabolic genes near late origins, supporting the notion that a gene dosage effect could be the results of the evolutionary pressure to cope with high energy costs during genomic duplication and that therefore, origin and metabolic gene sequences that co-localize might have evolved side by side. In essence, we learned that looking at the genes in origin proximity enables to predict the timely domain of activation.

Size control and genomic replication - bridging the gap

The most fundamental process in the biology of every living organism is reproduction, i.e. producing healthy descendants. Therefore, individuals are born, they grow into adults and then give birth. Herein, a considerable challenge is surviving long enough to grow and develop into fertile organisms to pass on the genomic information. Thus, reaching the fertile, adult age, or loosely phrased to grow-up, is an important part of the lifespan. Besides proliferation this means for the individual to reach a distinct size. Cook and Tyers (2007) write that

> "Despite the huge range of organism sizes across the biosphere, individuals

> within species are strikingly uniform, both in overall size and in the proportion and dimensions of organs within the body. This uniformity implies exquisite control of cell, organ and organism size mediated through elaborate coordination of cellular growth and proliferation."

Thus, size control is a universal feature occurring on many levels in biology whether in case of single cellular or higher, much more complex organisms. On the single cell level, growing and committing to cell division is deeply entangled with duplication of all cellular components, most importantly the DNA, to prepare faithful genetic inheritance for newborn cells. Neither contributing to the "[...] longstanding enigma of size control" (Cook and Tyers, 2007) nor to "fully comprehend the complexity of the chromosome replication process"(Hyrien and Goldar, 2009) in *S. cerevisiae* is a trivial goal. Due to the universality of the size homeostasis mechanism uncovered here (gating through metabolic power) and the high degree of conservation of the replication machinery, the studies in this model organism account for many life forms and must not be seen as isolated processes, but rather as one step towards the understanding of crucial cellular events, which, if deregulated, are often fatal and can lead to severe diseases in humans, such as cancer.

In this thesis, I present systems level studies of the growth rate, cell size and cell division linkage, complemented with a multi-level study of DNA replication to contribute to understanding the regulation of growth, reproduction and with it health. Using systems approaches enabled us to provide testable model predictions to guide future experiments and suggest follow-up studies for further theoretical analysis to increase the in-depth understanding of size control and genomic duplication. We learn that an integrative approach can yield insights on different levels of the biological system, but also that studying the system from different perspectives requires adapted levels of abstraction. Integration of different perspectives allows for a glance at the complexity of biological systems and I conclude that finding the appropriate level of abstraction can help in gaining knowledge and thus, contribute to fully understand biological systems, such as the cell cycle, cellular growth or genomic duplication in a larger context.

Appendix

Appendix A

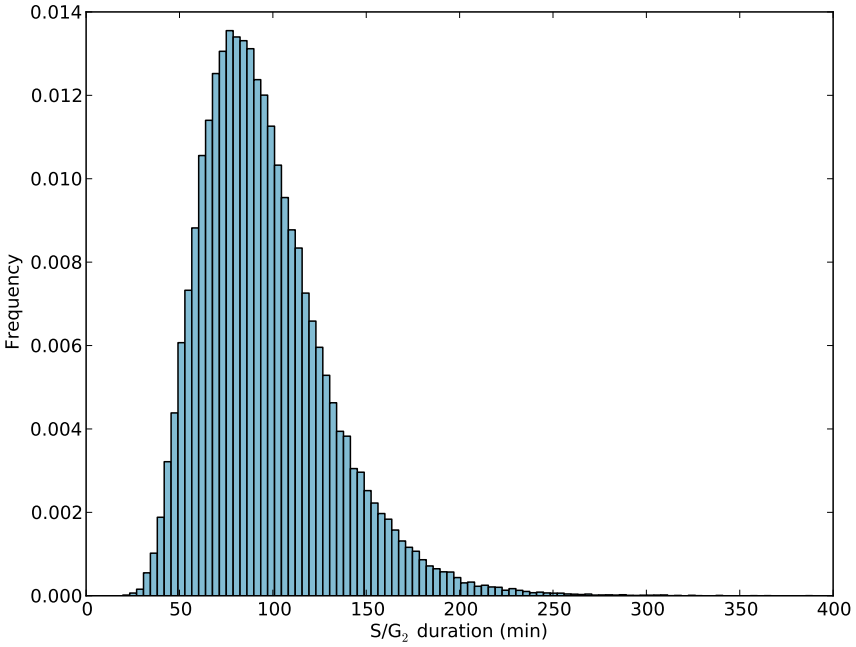


Figure A1: **Distribution of S/G_2 durations.** Presented are S/G_2 durations in minutes sampled from a normal distribution with mean 90 minutes and standard deviation around $\sim$ 10 % of the mean on a log scale.

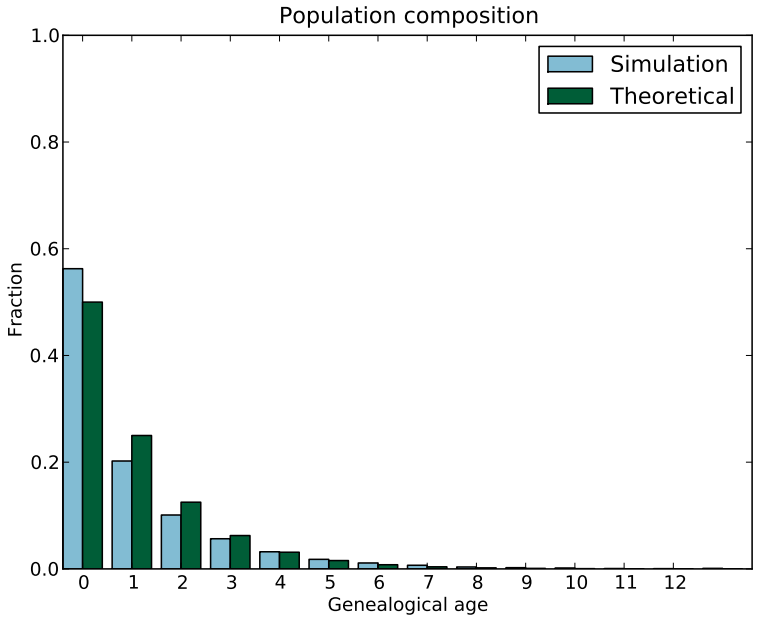


Figure A2: **The simulated culture is younger than an identical "ideal" culture.** In a theoretical cell culture without senescence and death, half of the cells will be newborn daughters, a quarter first generation mothers and so on. However, in the simulated culture we observe a shift from this distribution as the newborn daughters delay before dividing and hence there is a slight accumulation of daughter cells and a similar decrease in the fraction of mother cells as compared to the ideal culture. There appears to be a similar asymmetry within the mother line, with a relative accumulation of older mothers as compared to first and second generation mothers reflecting the decrease in G_1 duration with age.

Name	Specification	Initial value
$mCLN1,2$	cyclin Cln12-precurser (mRNA)	0
Cln12	Cdc28 is always available and therefore, not explicitly modeled but implied. Cln12 represents the active kinase complex.	0
B^R	internal biomass	25
B^{Am}	structural biomass (mother)	8.5
B^{Ad}	structural biomass (daughter)	0
mB^R	B^R-precurser	1
mB^A	B^A-precurser	1

Table A1: **List of model species and initial values.**

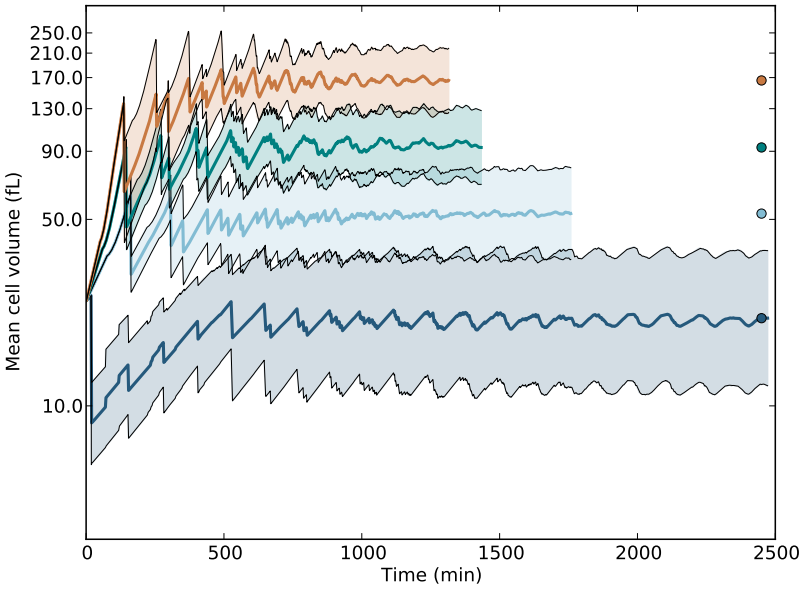
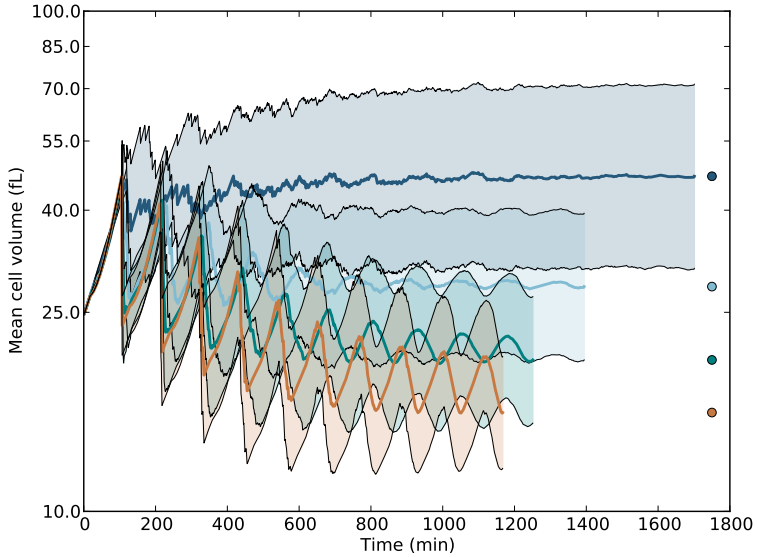
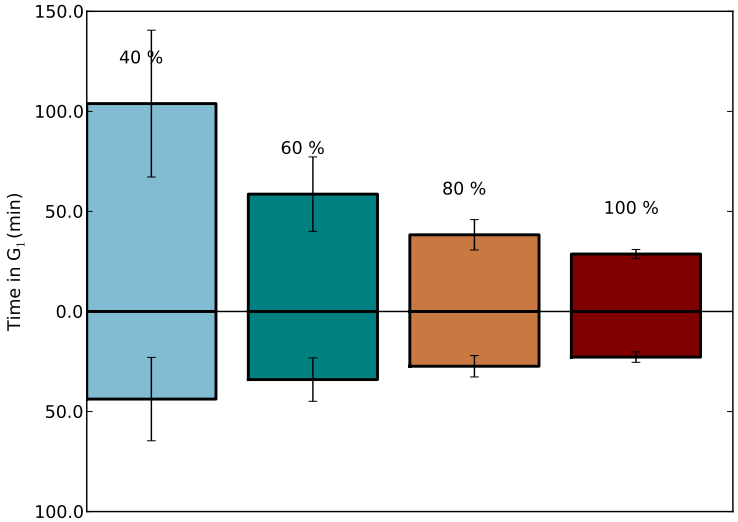


Figure A3: **Deterministic modeling yields the same qualitative effect but cells take longer to lose synchrony.** The model was adjusted to give deterministic expression of *mCLN1/2* during the G_1 phase corresponding to the expected value of the stochastic model (0.4 instead of 40% chance of 1). The simulations shown correspond to four different growth rates as in Figure 2.9. While the cells take longer to desynchronise, the qualitative behaviour is unchanged. Hence, the stochastic simulation has the advantage of approaching an asynchronous steady state faster and hence with much less cells (and computational burden).



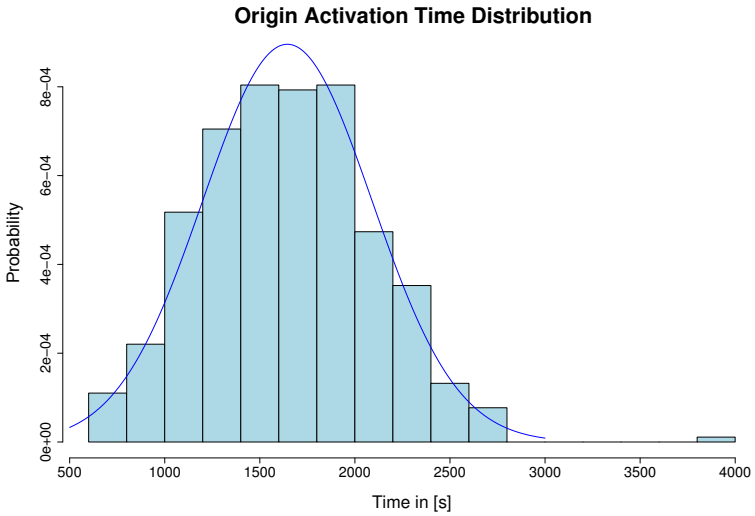
(a)



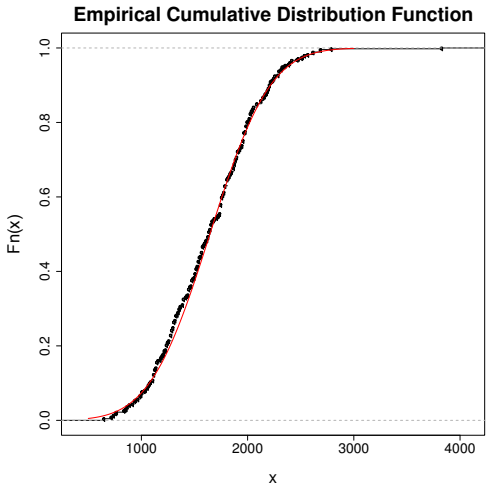
(b)

Figure A4: **Cln overactivation leads to shorter G_1 duration, smaller cells and higher growth rate.** The effect of *CLN3* overexpression, increased growth rate but decreased size, has been pointed out as a paradox (Hall et al., 1998; Barberis et al., 2007). Here, we approach this issue by incremental increase in *CLN1/2* expression as the model lacks Cln3. The probability for expression was raised from 40% (blue) to 60% (green), 80% (orange) or 100% (red). Increasing production of Cln1/2 leads to smaller cell size (a), slower decay of synchrony (a) and decreased time in G_1 (b). The growth rate, calculated as 1/generation time in hours, is also increasing from 0.37 to 0.43, 0.47 or 0.48, respectively.

Appendix B



(a)



(b)

Figure B5: **Distribution of origin firing times.** The distribution was approximated by a normal distribution (blue line) (a) and the cumulative distribution was calculated to show similarities of firing time and normal distribution (b).

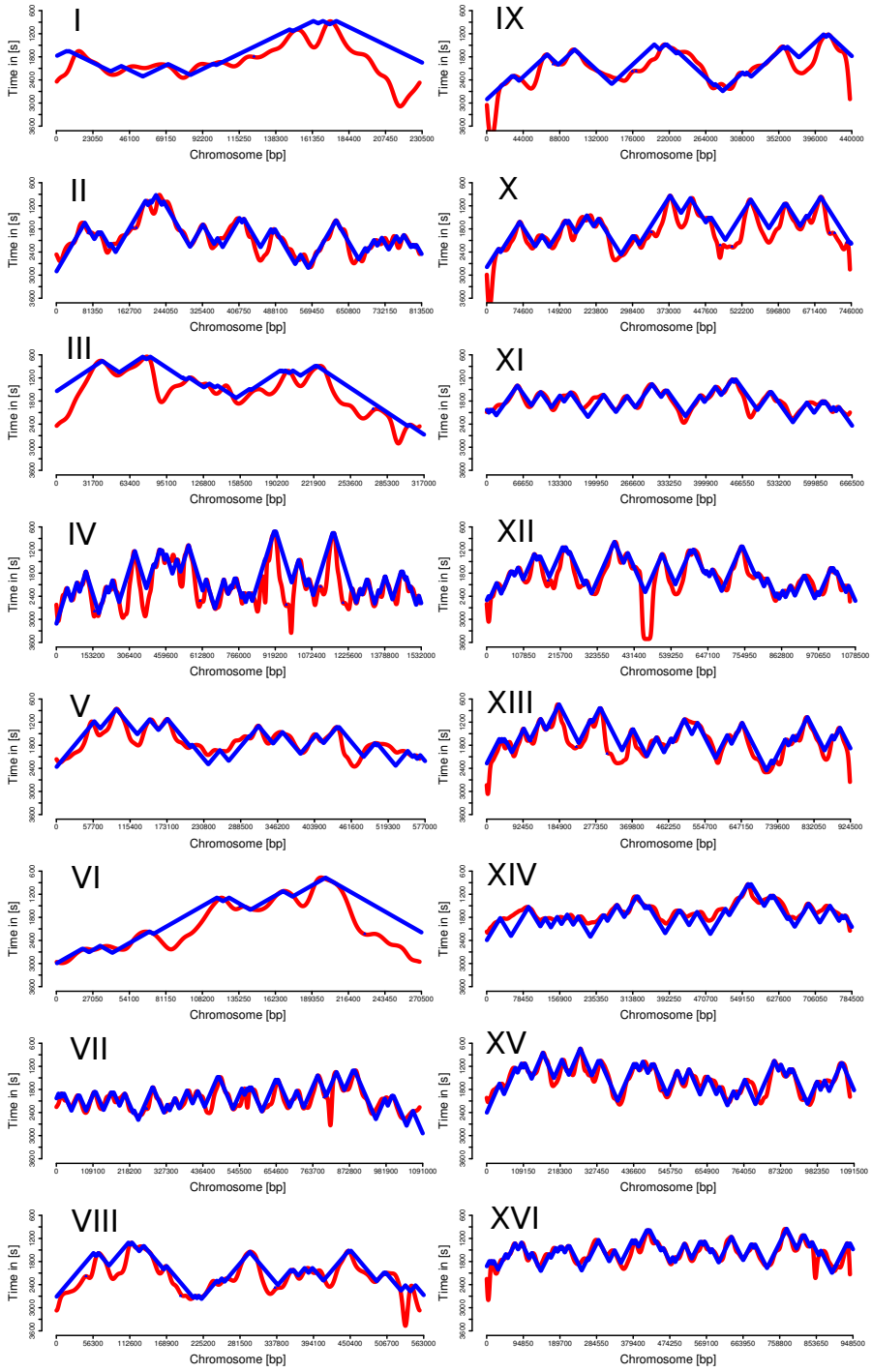


Figure B6: **Experimental and simulated replication profiles for chromosomes I-XVI.** Red curves are recalculated according to the microarray-based heavy:light data from Raghuraman et al. (2001) and blue curves represent the simulated profiles. The replication time in seconds is plotted as a function of chromosome coordinate in base pairs (bp). Single figures can also be found in the electronic supplementary material of Spiesser et al. (2009).

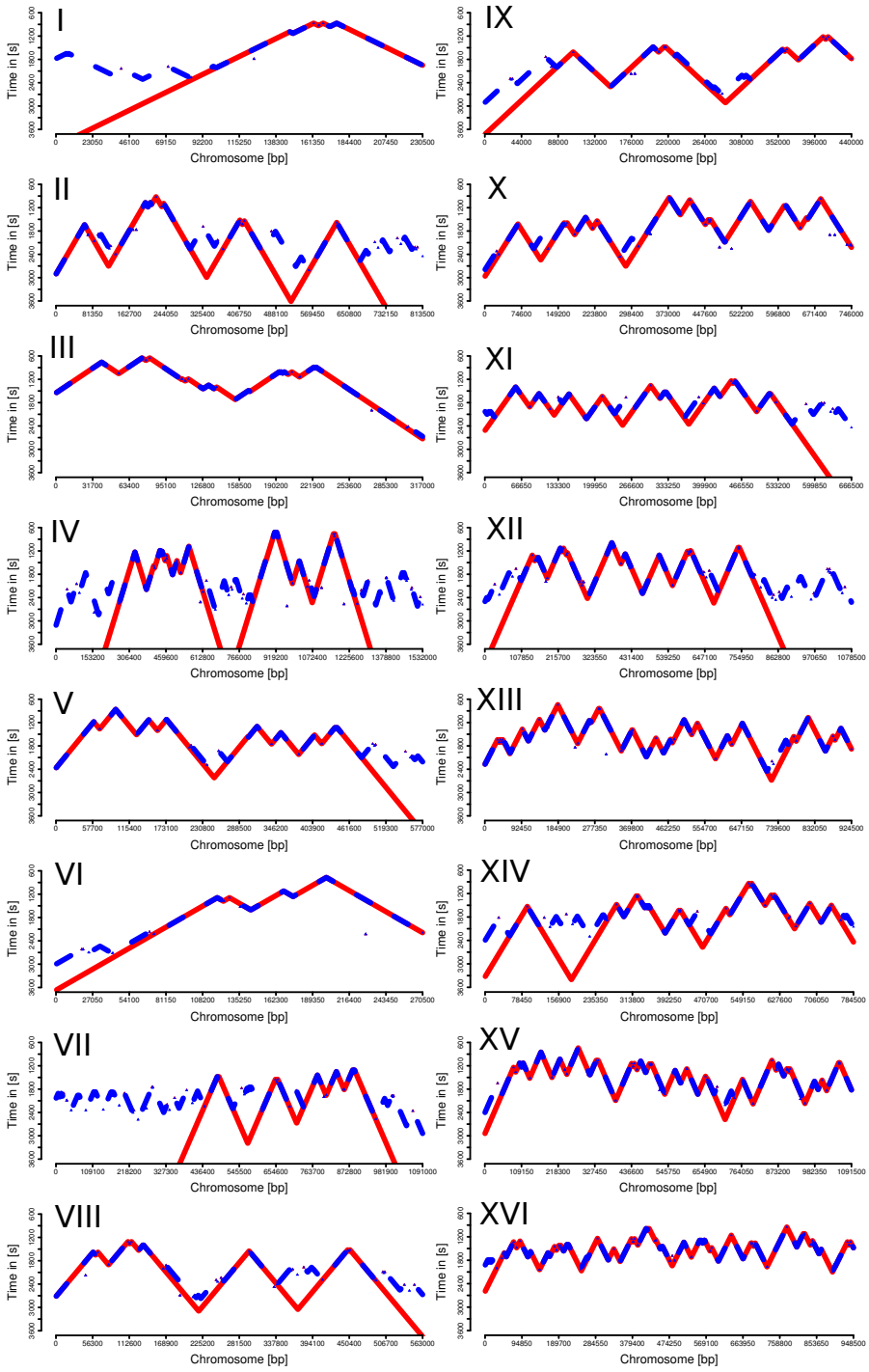


Figure B7: **Simulated replication profiles in wild type and *clb5*△ background for chromosomes I-XVI.** The dotted blue line represents the simulated profile for wild type cells and the red one represents the computed profile for the *clb5*△ mutant. Single figures can also be found in the electronic supplementary material of Spiesser et al. (2009).

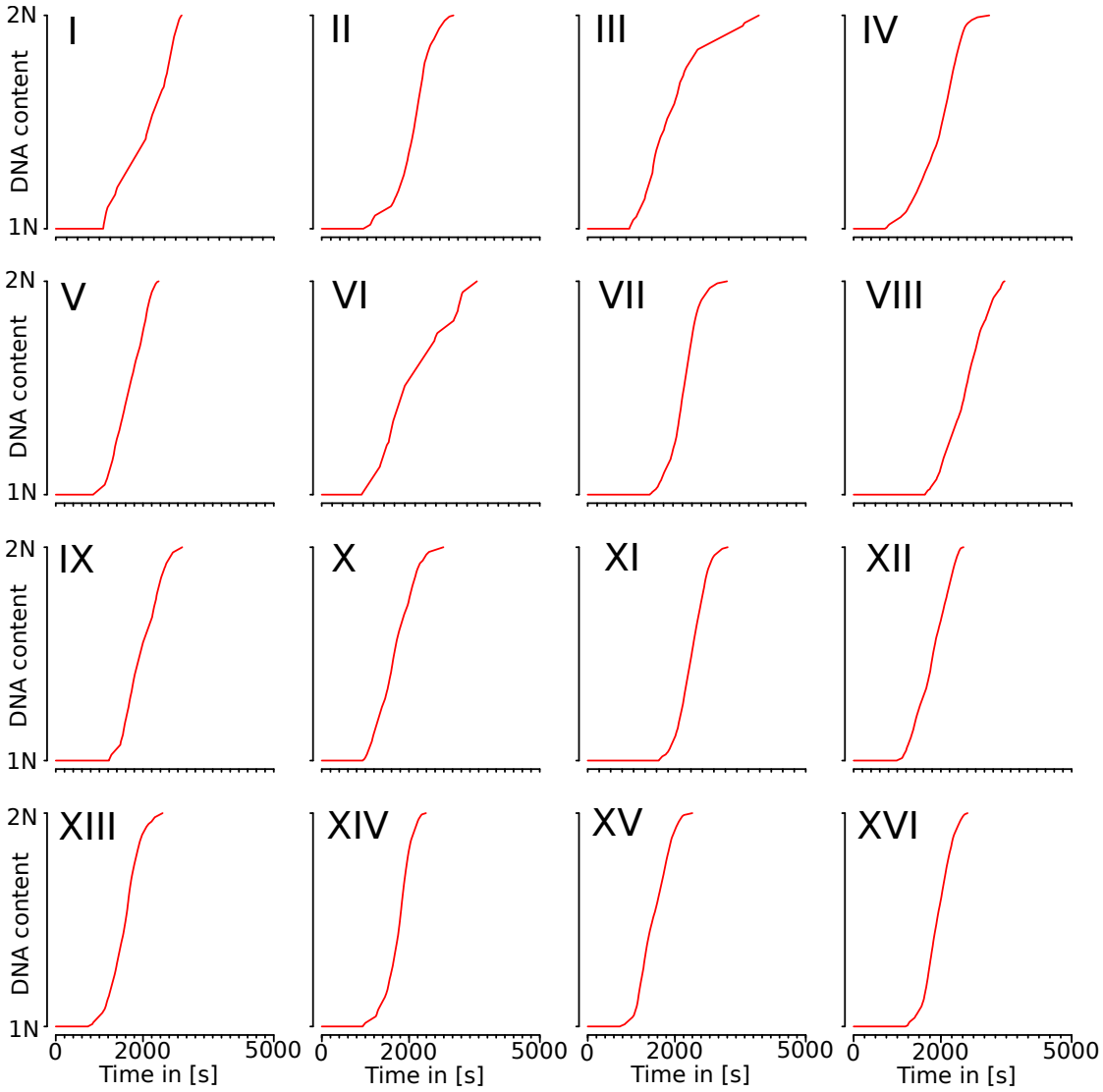


Figure B8: **Simulated replication kinetics for wild type cells for chromosomes I-XVI.** The simulations show the increase of DNA content over time. Single figures can also be found in the electronic supplementary material of Spiesser et al. (2009).

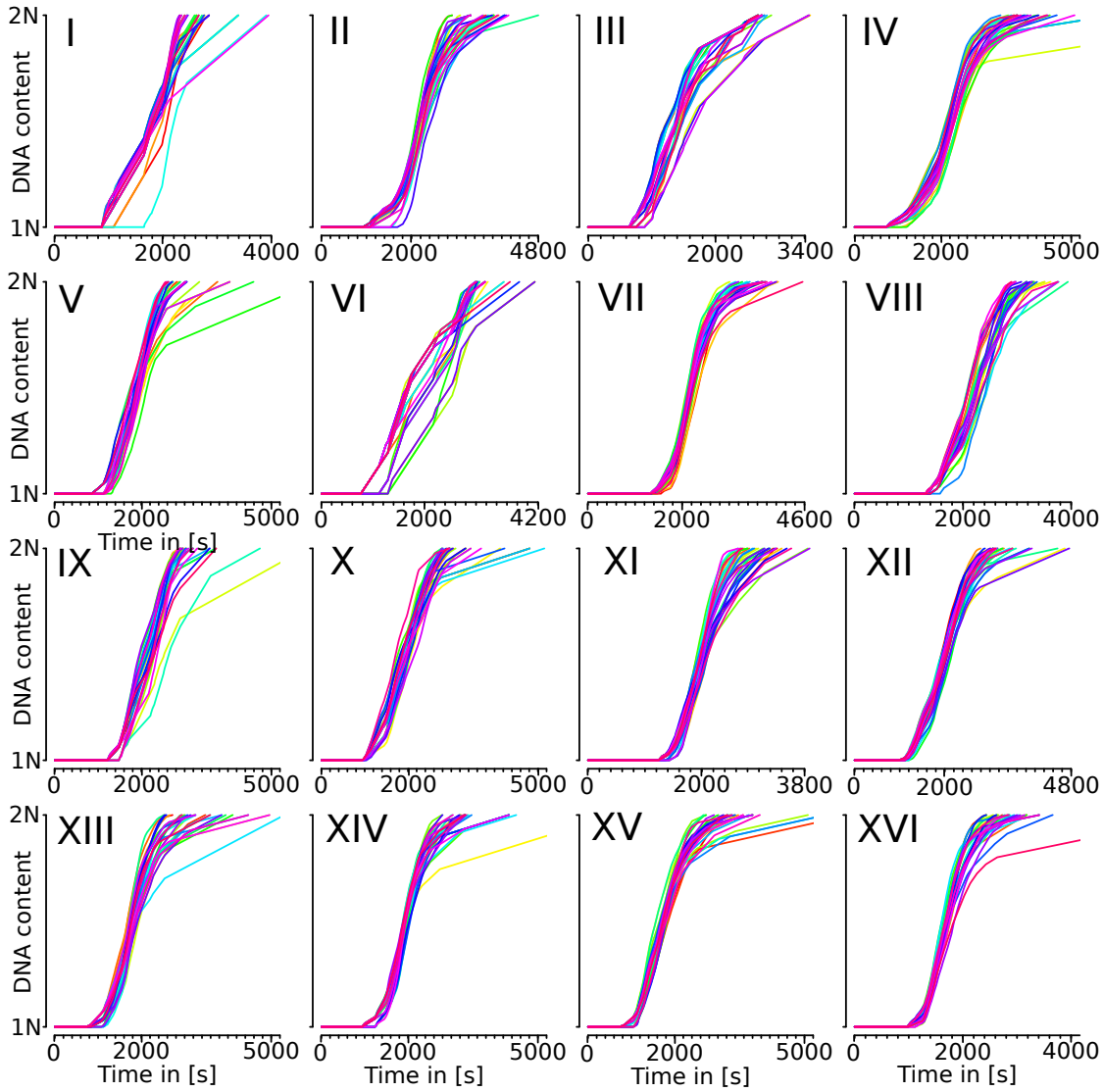


Figure B9: **Simulated replication kinetics for perturbed cells for chromosomes I-XVI.** The simulations have been performed with 30 reduced sets of replication origins derived from random deletion of 50% of the original origins. Single figures can also be found in the electronic supplementary material of Spiesser et al. (2009).

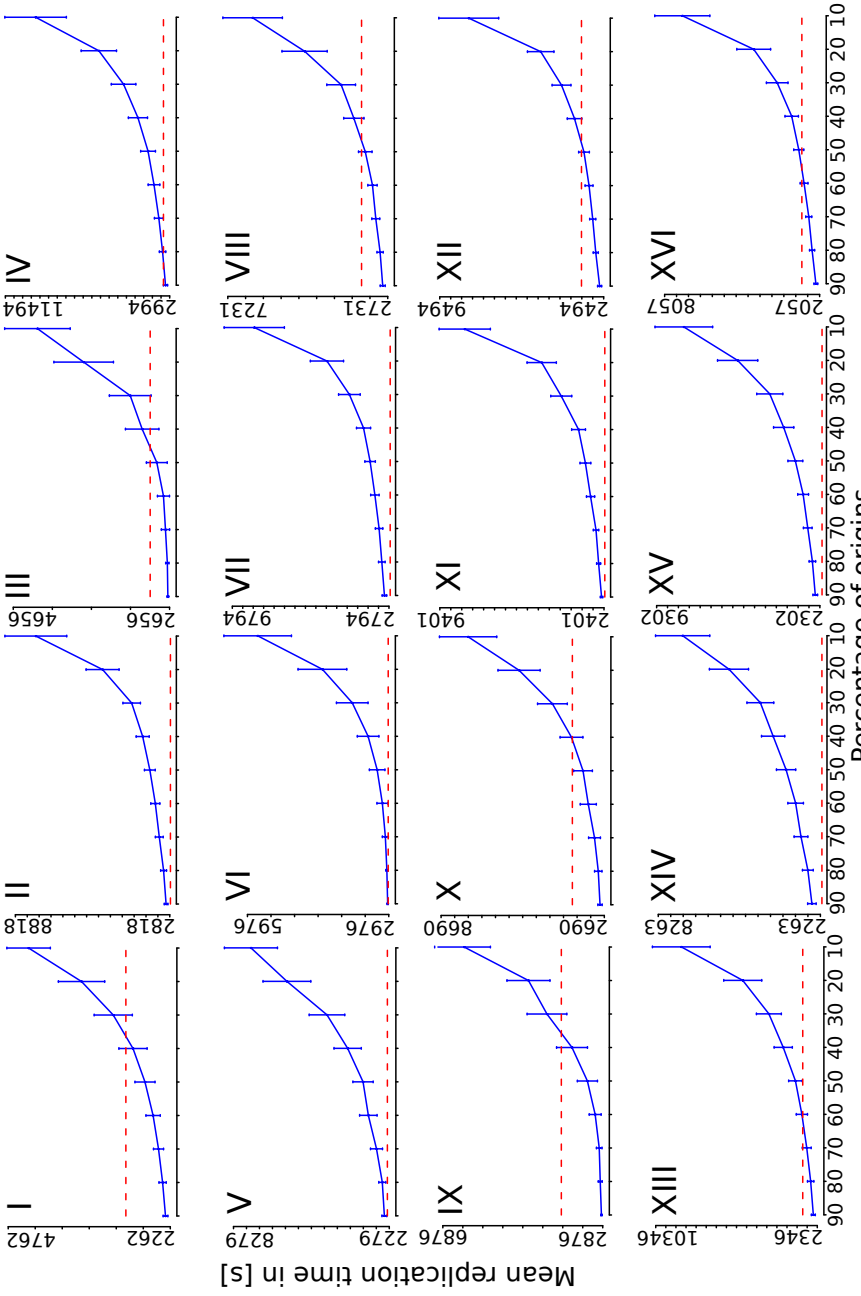


Figure B10: **Mean replication time for chromosomes I-XVI.** Blue lines represent curves for descending percentages of replication origins (from 90% to 10%). Error bars show standard deviation of 10,000 simulations. Dashed red lines indicate experimental replication times for each chromosome (Raghuraman et al., 2001). Single figures can also be found in the electronic supplementary material of Spiesser et al. (2009).

Appendix C

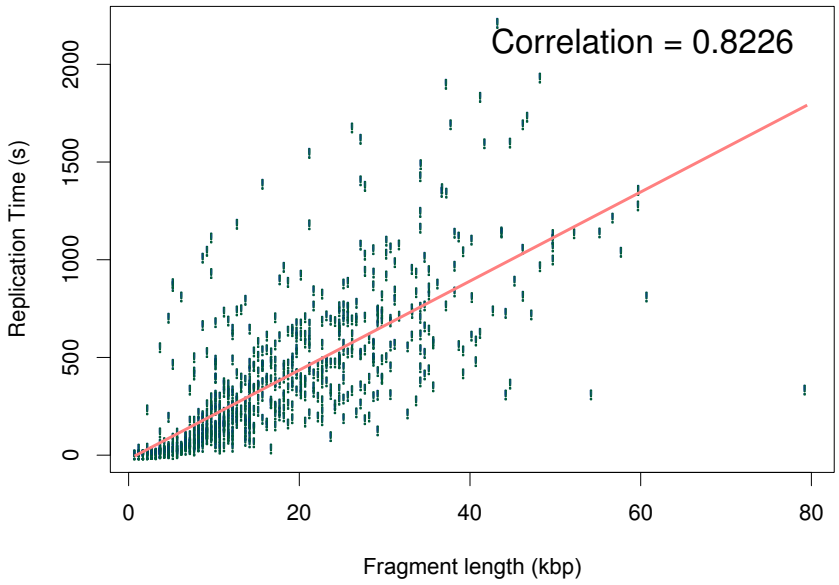


Figure C11: **Dependence of replication times on the lengths of the DNA templates.** In the experimental data (Raghuraman et al., 2001) a significant correlation between the length of the replicated DNA template and the replication time (~ 0.82, Spearman-Rank Correlation) is observed.

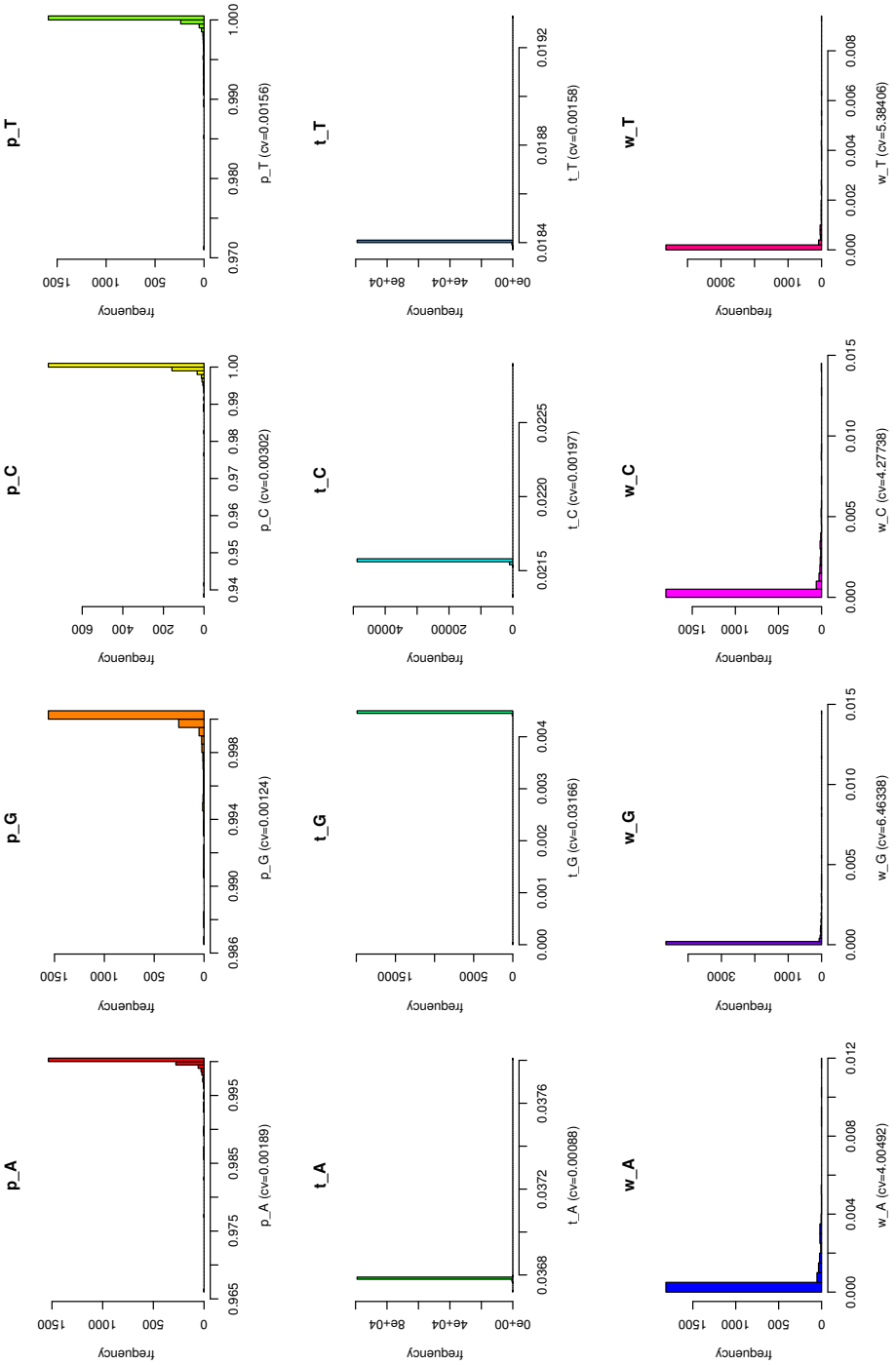


Figure C12: **Estimated parameters for *model 1*.** Histograms for the 12 parameters from 1000 independent optimization runs with uniformly distributed initial values. CV denotes coefficient of variation.

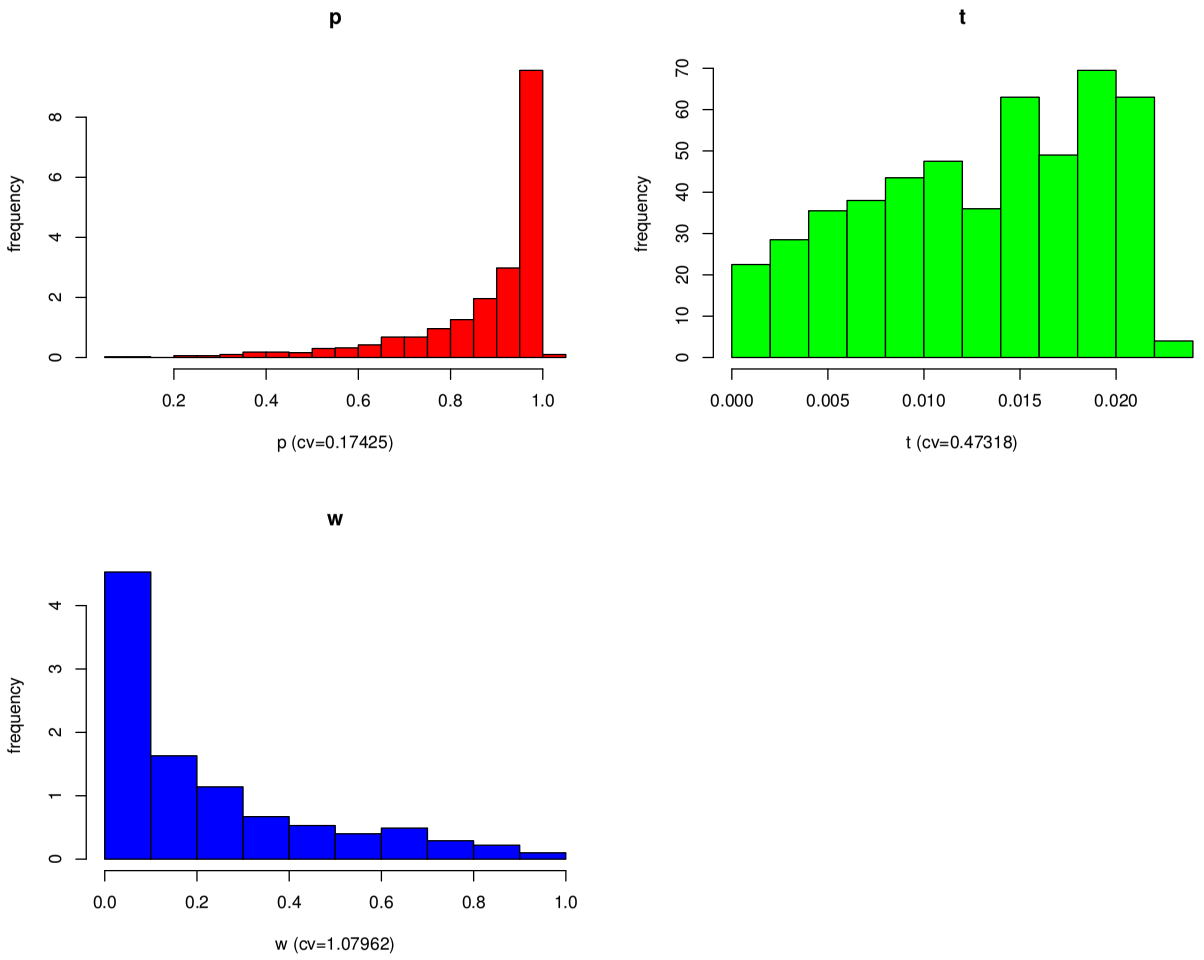


Figure C13: **Estimated parameters for *model 2*.** Histograms for the 3 parameters from 1000 independent optimization runs with uniformly distributed initial values. CV denotes coefficient of variation.

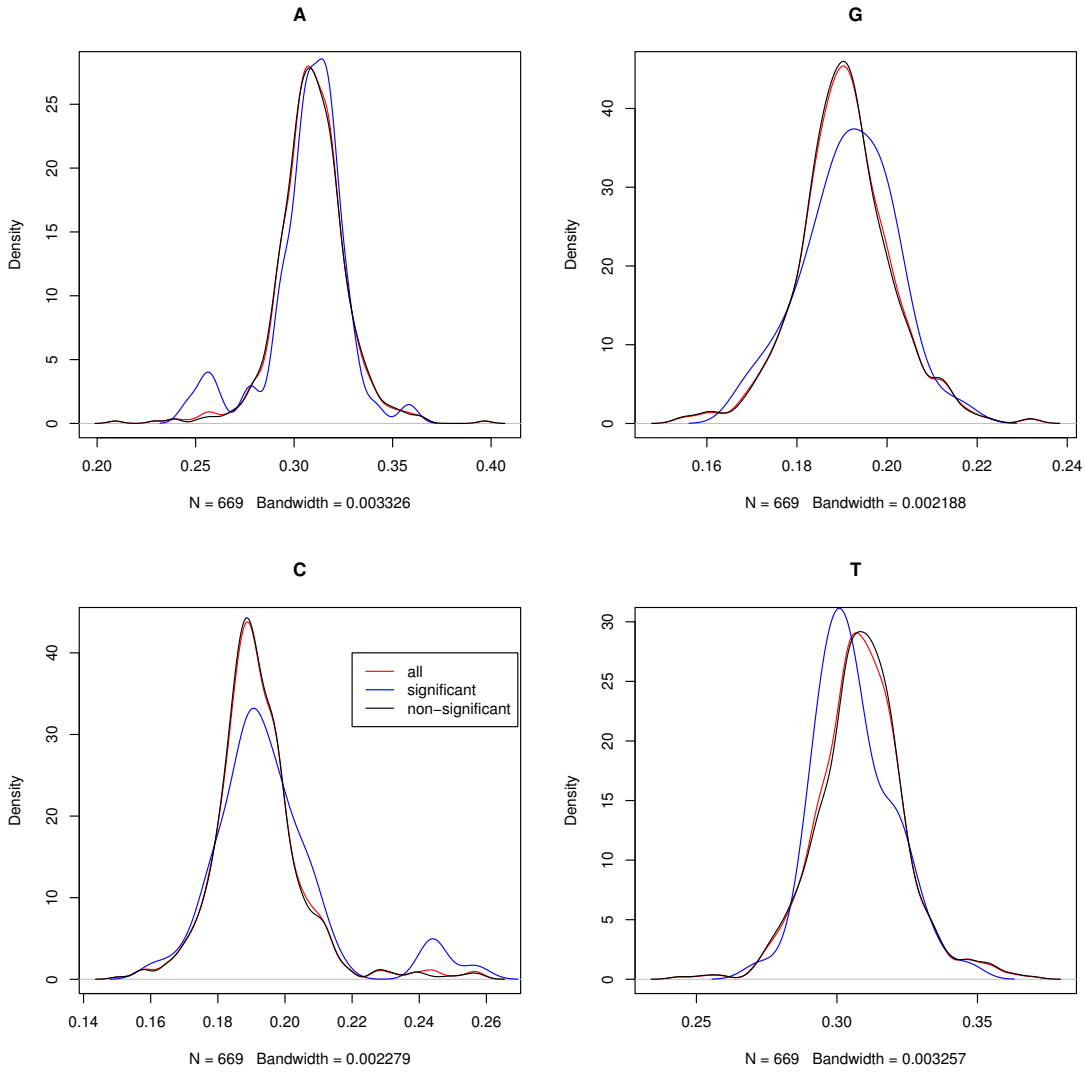


Figure C14: **Single nucleotide density estimates** in significant (blue), non-significant (black) or all (red) regions. Estimation of underlying distribution was calculated using a non-parametric estimator, as defined in equation 5.2.

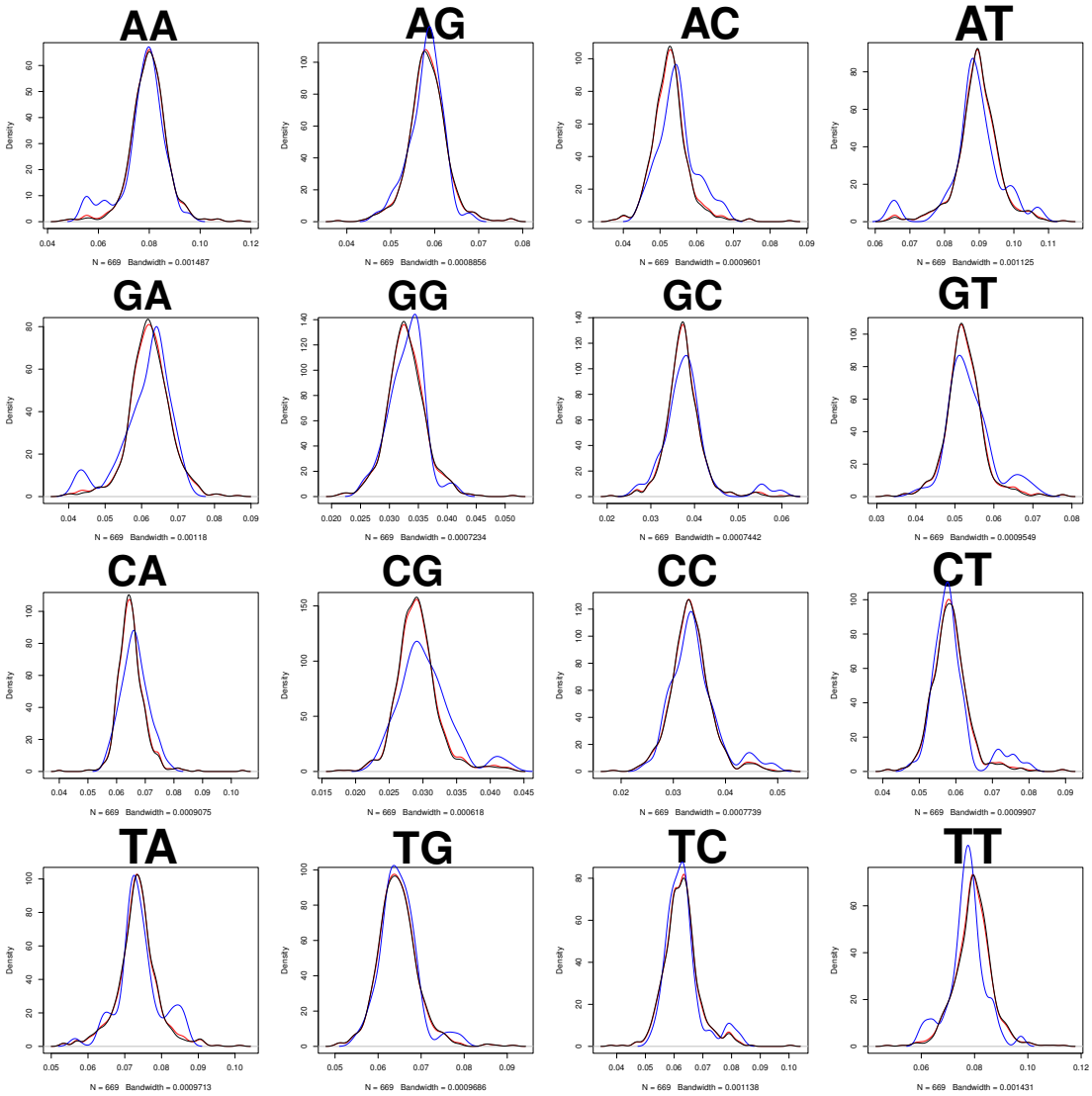


Figure C15: **Pair wise nucleotide density estimates** in significant (blue), non-significant (black) or all (red) regions. Estimation of underlying distribution was calculated using a non-parametric estimator, as defined in equation 5.2.

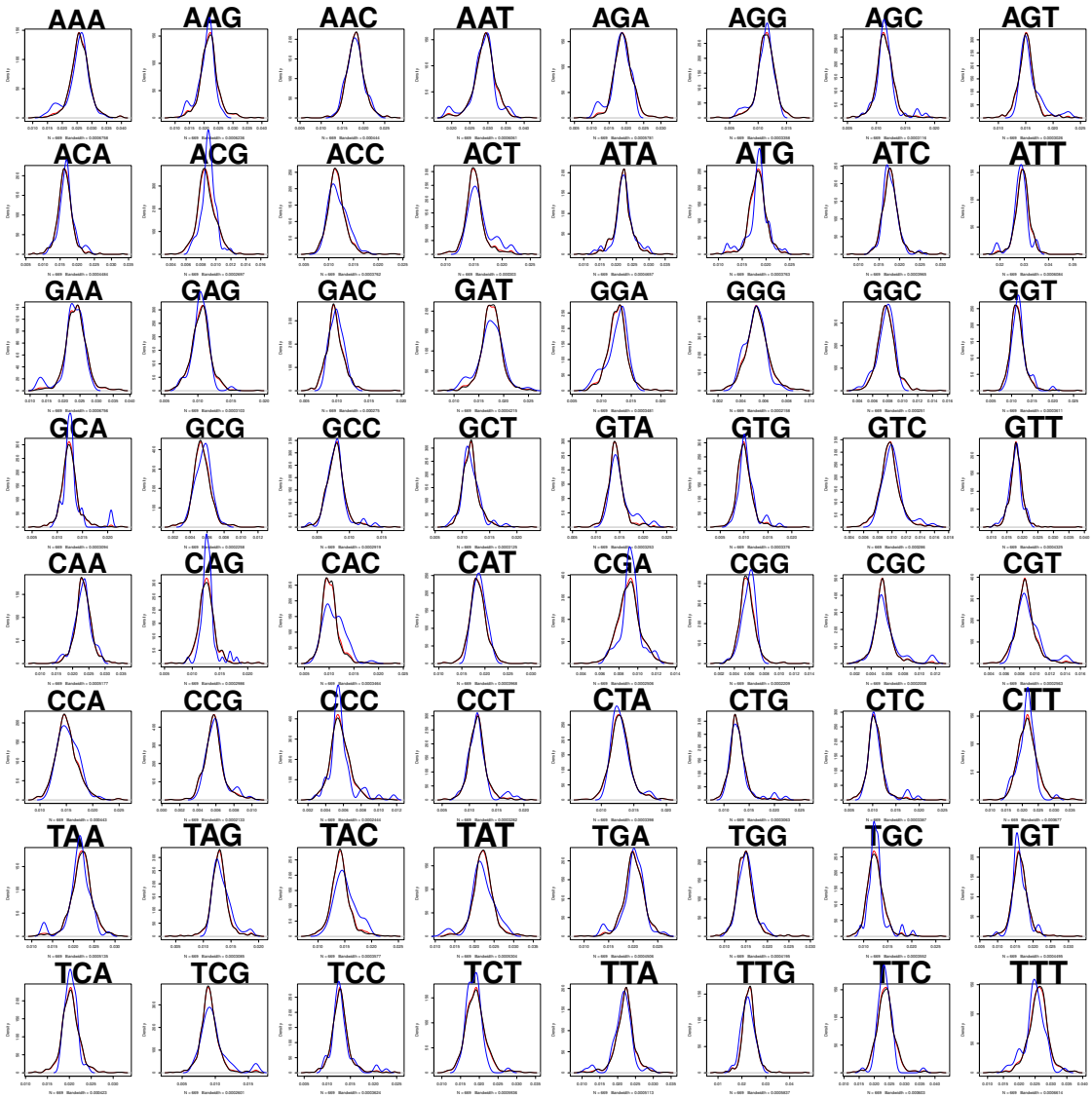
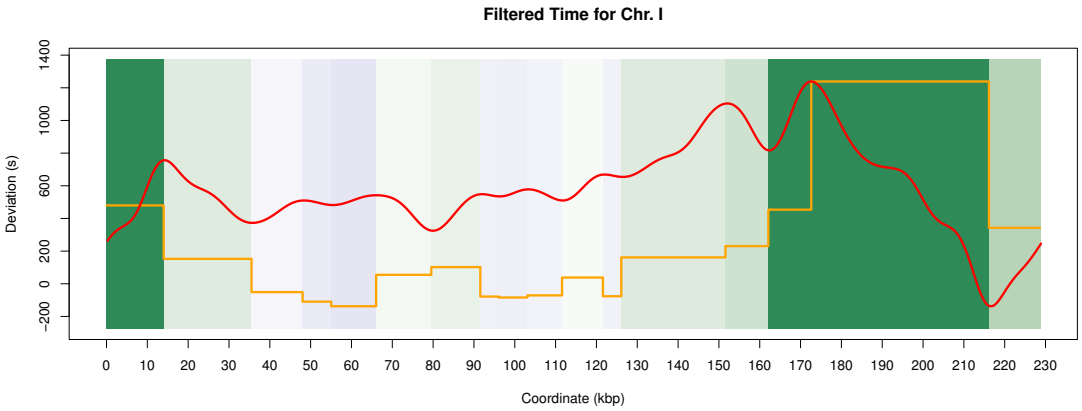
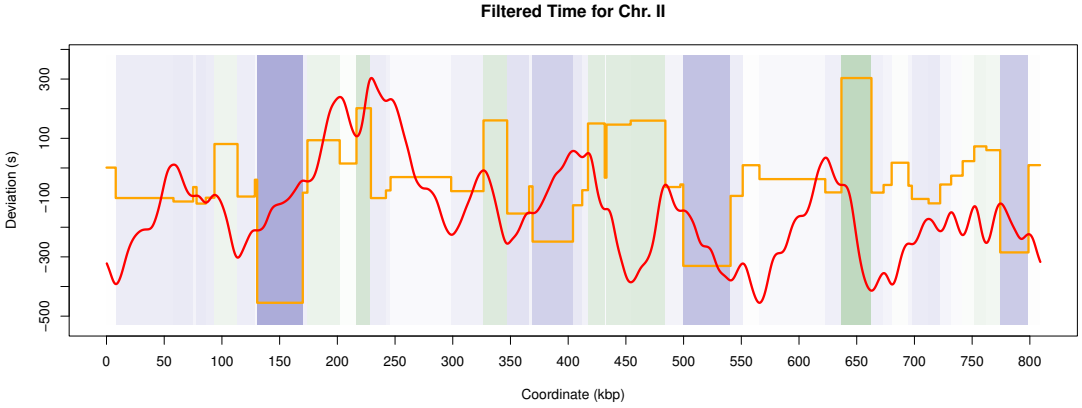


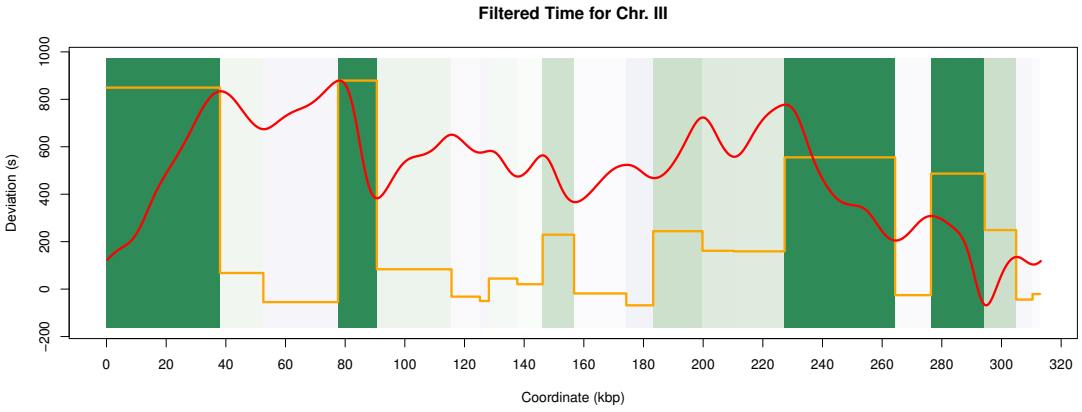
Figure C16: **Triple wise nucleotide density estimates** in significant (blue), non-significant (black) or all (red) regions. Estimation of underlying distribution was calculated using a non-parametric estimator, as defined in equation 5.2.


Filtered Time for Chr. I
Deviation (s)
−200
0
200
600
1000
1400
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
Coordinate (kbp)

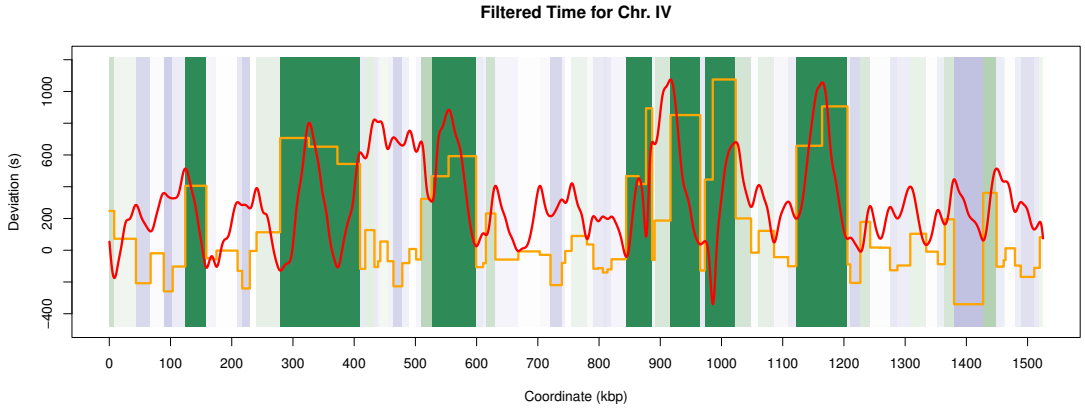
(a)


Filtered Time for Chr. II
Deviation (s)
−500
−300
−100
100
300
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
Coordinate (kbp)

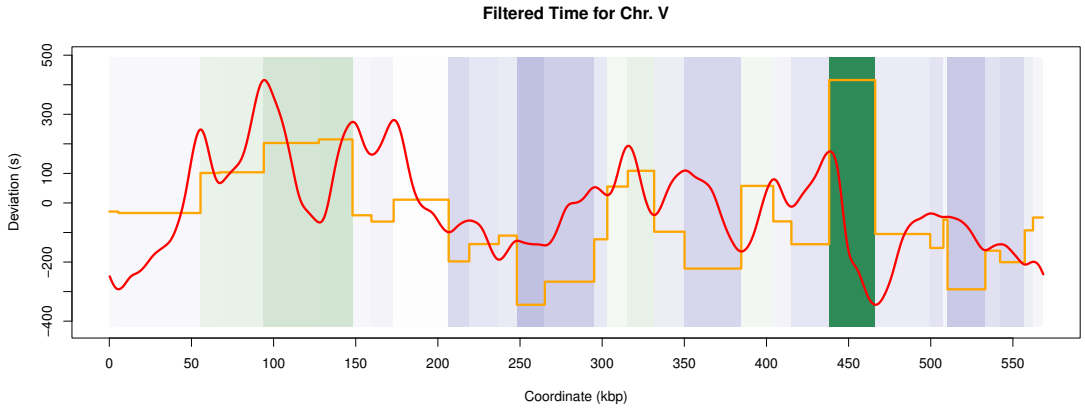
(b)


Filtered Time for Chr. III
Deviation (s)
−200
0
200
400
600
800
1000
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
Coordinate (kbp)

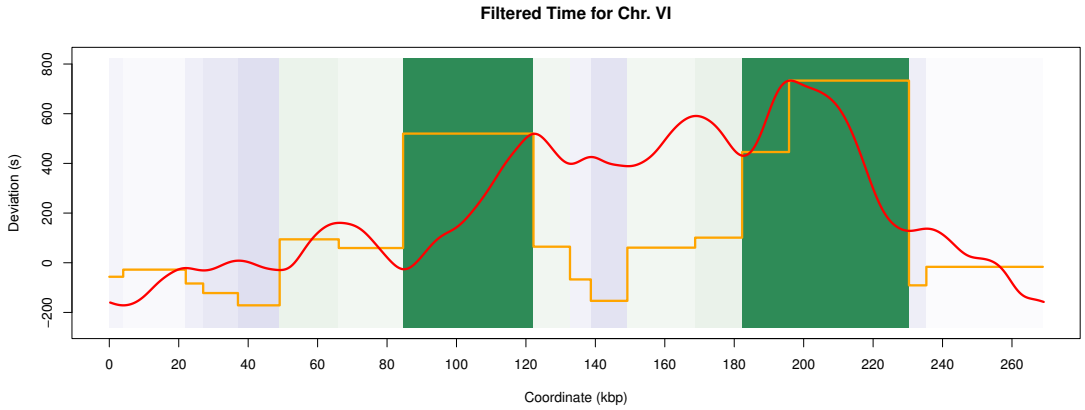
(c)


Filtered Time for Chr. IV
Deviation (s)
−400
0
200
600
1000
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
Coordinate (kbp)

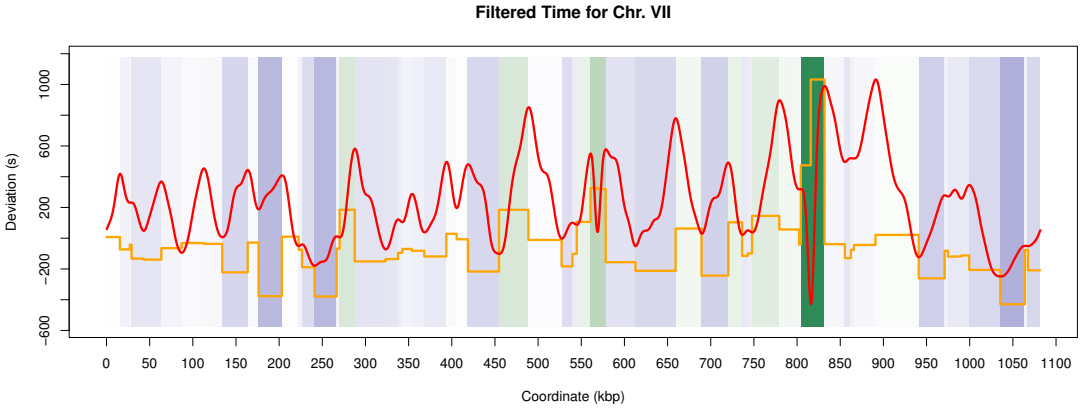
(d)


Filtered Time for Chr. V
Deviation (s)
−400
−200
0
100
300
500
0
50
100
150
200
250
300
350
400
450
500
550
Coordinate (kbp)

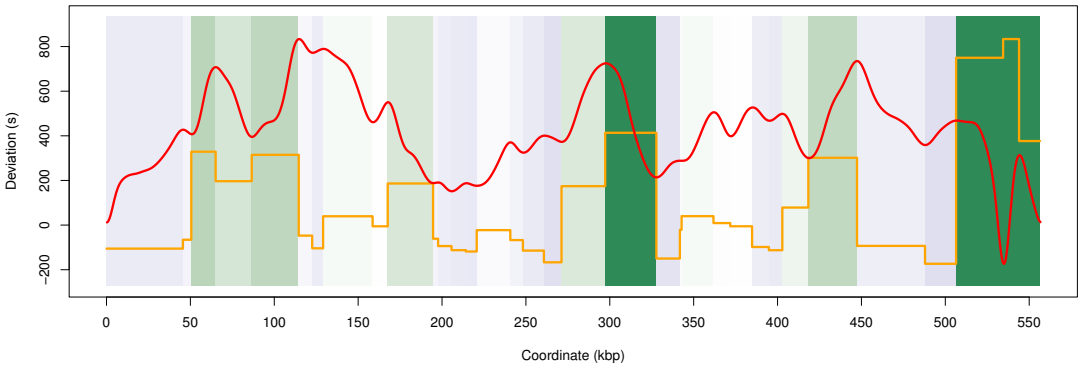
(e)


Filtered Time for Chr. VI
Deviation (s)
−200
0
200
400
600
800
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Coordinate (kbp)

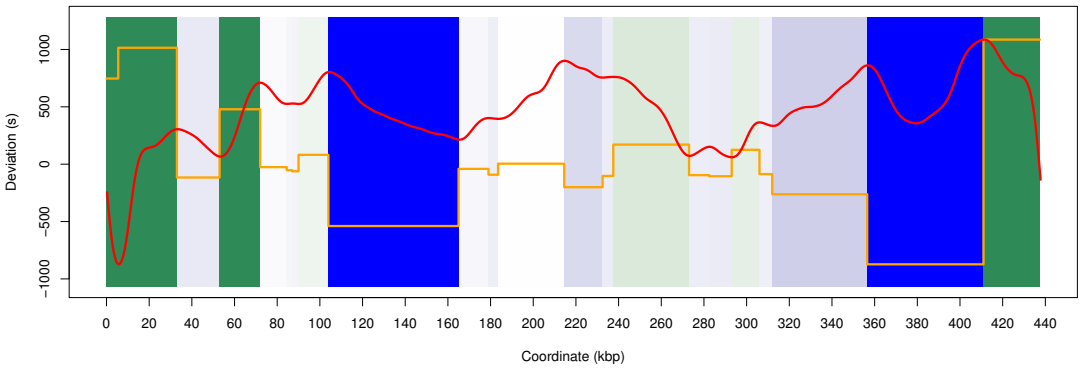
(f)


Filtered Time for Chr. VII
Deviation (s)
−600
−200
200
600
1000
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
Coordinate (kbp)

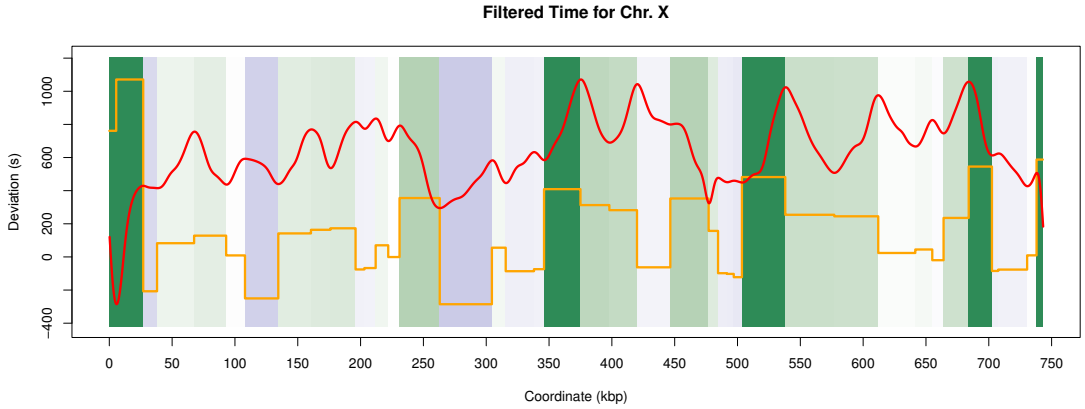
(g)


Filtered Time for Chr. VIII
Deviation (s)
−200
0
200
400
600
800
0
50
100
150
200
250
300
350
400
450
500
550
Coordinate (kbp)

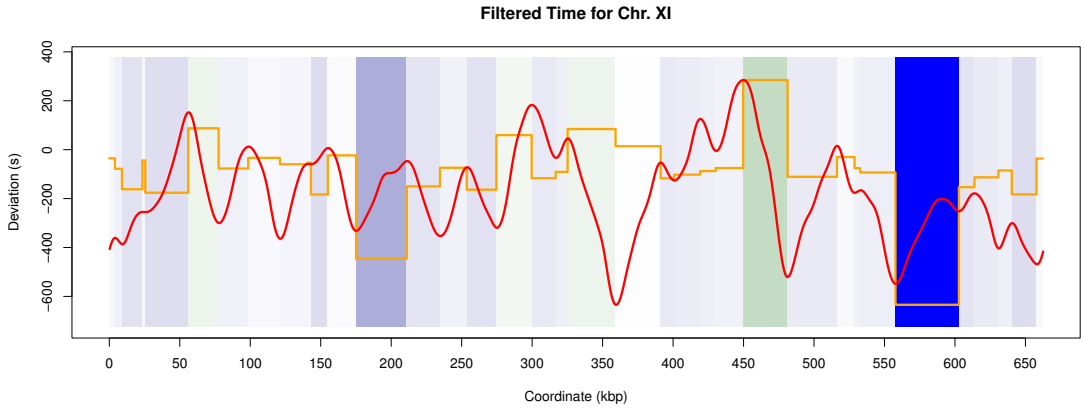
(h)


Filtered Time for Chr. IX
Deviation (s)
−1000
−500
0
500
1000
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
420
440
Coordinate (kbp)

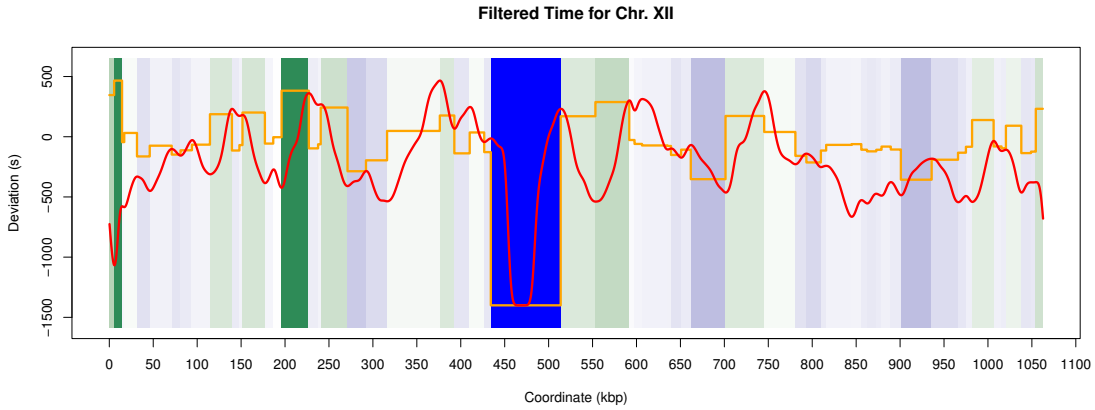
(i)


Filtered Time for Chr. X
Deviation (s)
−400
0
200
600
1000
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
Coordinate (kbp)

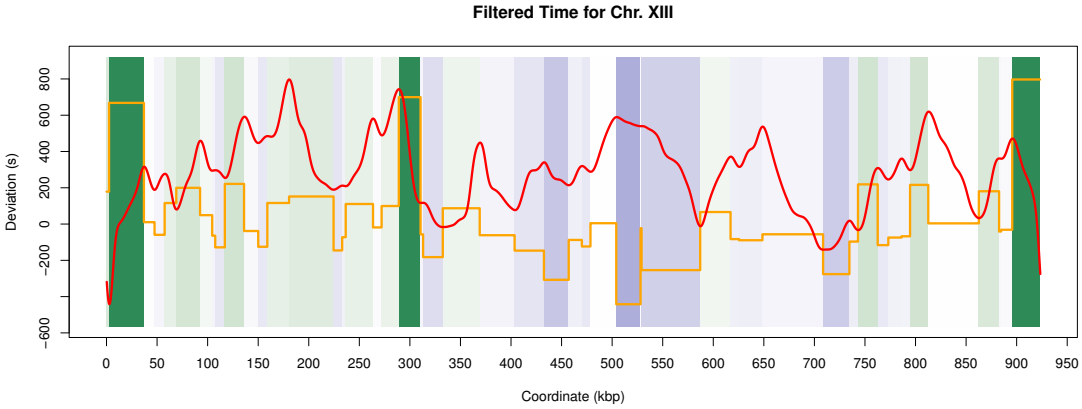
(j)


Filtered Time for Chr. XI
Deviation (s)
−600
−400
−200
0
200
400
0
50
100
150
200
250
300
350
400
450
500
550
600
650
Coordinate (kbp)

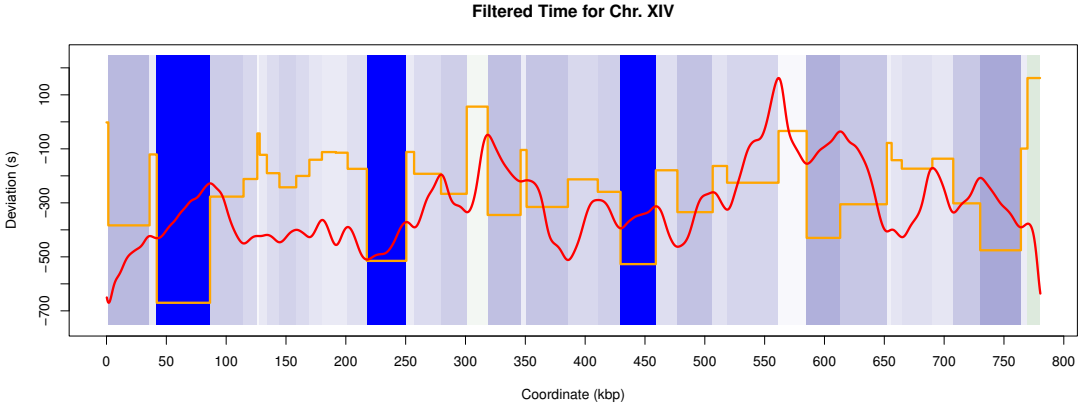
(k)


Filtered Time for Chr. XII
Deviation (s)
−1500
−1000
−500
0
500
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
Coordinate (kbp)

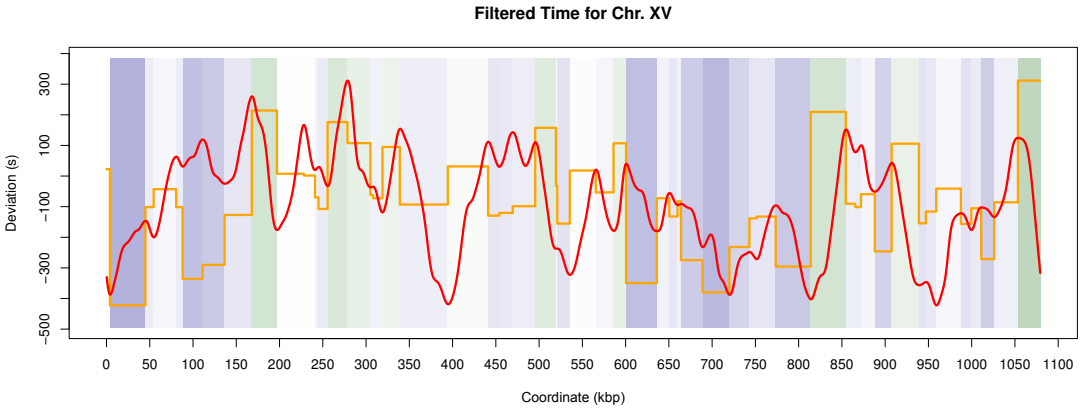
(l)


Filtered Time for Chr. XIII
Deviation (s)
−600
−200
0
200
400
600
800
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
Coordinate (kbp)

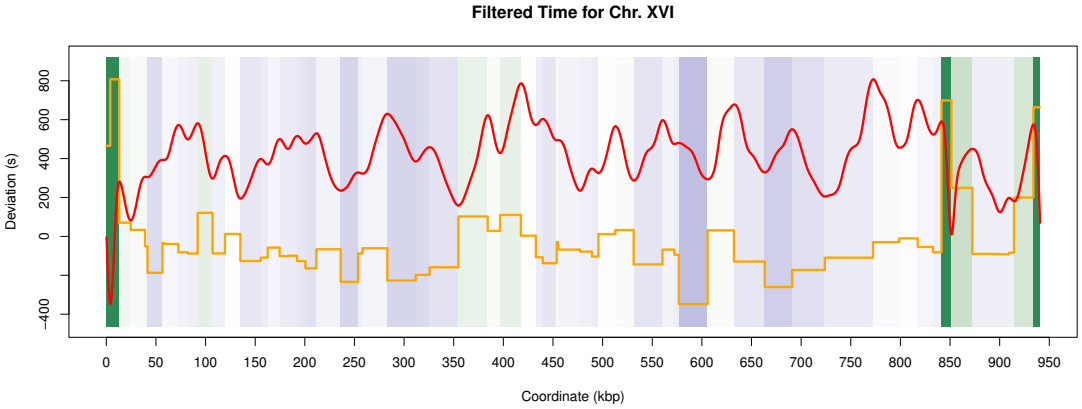
(m)


Filtered Time for Chr. XIV
Deviation (s)
−700
−500
−300
−100
100
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
Coordinate (kbp)

(n)


Filtered Time for Chr. XV
Deviation (s)
−500
−300
−100
100
300
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
1050
1100
Coordinate (kbp)

(o)



(p)

Figure C17: **Filtered times and experimental replication profiles** mapped onto the 16 chromosomes of budding yeast. The filtered times mapped onto the locations of their corresponding DNA segments are shown. The shadings correspond to the ones used in Figure 4.4. The orange line denotes the actual filtered time in seconds and the red line shows the replication profile from Raghuraman et al. (2001).

Bibliography

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 9 edition, 1972. ISBN 9780486612720.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

L. Alberghina, R. L. Rossi, L. Querin, V. Wanke, and M. Vanoni. A cell sizer network involving Cln3 and Far1 controls entrance into S phase in the mitotic cycle of budding yeast. *The Journal of Cell Biology*, 167(3):433–43, Nov. 2004.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science, 5 edition, 2007. ISBN 9780815341055.

M. Aldea, E. Gari, and N. Colomina. Control of cell cycle and cell growth by molecular chaperones. *Cell Cycle*, 6(21):2599, 2007.

A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22 (13):1600–1607, July 2006.

U. Alon. How To Choose a Good Scientific Problem. *Molecular Cell*, 35(6):726–728, 2009.

G. M. Alvino, D. Collingwood, J. M. Murphy, J. Delrow, B. J. Brewer, and M. K. Raghuraman. Replication in hydroxyurea: it's a matter of time. *Molecular and Cellular Biology*, 27(18):6396–406, Sept. 2007.

F. Antequera. Genomic specification and epigenetic regulation of eukaryotic DNA replication origins. *The EMBO Journal*, 23(22):4365–70, Nov. 2004.

O. M. Aparicio, A. M. Stout, and S. P. Bell. Differential assembly of Cdc45p and DNA polymerases at early and late origins of DNA replication. *Proceedings of the National Academy of Sciences USA*, 96(16):9130–9135, Aug. 1999.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

M. Barberis and E. Klipp. Insights into the network controlling the G(1)/S transition in budding yeast. *Genome Informatics*, 18:85–99, 2007.

M. Barberis, E. Klipp, M. Vanoni, and L. Alberghina. Cell size at S phase initiation: an emergent property of the G1/S network. *PLoS Computational Biology*, 3(4):e64, Apr. 2007.

M. Barberis, T. W. Spiesser, and E. Klipp. Replication origins and timing of temporal replication in budding yeast: how to solve the conundrum. *Current Genomics*, 11(3): 199–211, May 2010.

J. P. Barford and R. J. Hall. Estimation of the length of cell cycle phases from asynchronous cultures of Saccharomyces cerevisiae. *Experimental Cell Research*, 102(2): 276–284, Oct. 1976.

D. Barik, W. T. Baumann, M. R. Paul, B. Novak, and J. J. Tyson. A model of yeast cell-cycle regulation based on multisite phosphorylation. *Molecular Systems Biology*, 6:405, Aug. 2010.

J. Bechhoefer and B. Marshall. How Xenopus Laevis Replicates DNA Reliably even though Its Origins of Replication are Located and Initiated Stochastically. *Physical Review Letters*, 98(9):1–4, Feb. 2007.

S. P. Bell and A. Dutta. DNA replication in eukaryotic cells. *Annual Review of Biochemistry*, 71:333–374, 2002.

N. M. Berbenetz, C. Nislow, and G. W. Brown. Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure. *PLoS Genetics*, 6 (9):13, Sept. 2010.

K. A. Bernstein, F. Bleichert, J. M. Bean, F. R. Cross, and S. J. Baserga. Ribosome biogenesis is sensed at the Start cell cycle checkpoint. *Molecular Biology of the Cell*, 18(3):953–964, Mar. 2007.

A.-K. Bielinsky. Replication origins: why do we need so many? *Cell Cycle*, 2(4):307–309, 2003.

J. Bloom and F. R. Cross. Multiple levels of cyclin specificity in cell-cycle control. *Nature Reviews Molecular Cell Biology*, 8(2):149–160, Feb. 2007.

E. M. Boczko, T. G. Cooper, T. Gedeon, K. Mischaikow, D. G. Murdock, S. Pratap, and K. S. Wells. Structure theorems and the dynamics of nitrogen catabolite repression in yeast. *Proceedings of the National Academy of Sciences USA*, 102(16):5647–5652, Apr. 2005.

G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley, New York, 1987. ISBN 0471810339.

A. M. Breier, S. Chatterji, and N. R. Cozzarelli. Prediction of Saccharomyces cerevisiae replication origins. *Genome Biology*, 5(4):R22, 2004.

B. J. Brewer, E. Chlebowicz-Sledziewska, and W. L. Fangman. Cell cycle phases in the unequal mother/daughter cell cycles of Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 4(11):2529–31, Nov. 1984.

F. J. Bruggeman, N. Blüthgen, and H. V. Westerhoff. Noise management by molecular networks. *PLoS Computational Biology*, 5(9):e1000506, Sept. 2009.

A. Brümmer, C. Salazar, V. Zinzalla, L. Alberghina, and T. Höfer. Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. *PLoS Computational Biology*, 6(5):e1000783, Jan. 2010.

P. Burgers. Polymerase dynamics at the eukaryotic DNA replication fork. *Journal of Biological Chemistry*, 284(7):4041, Feb. 2009.

R. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208, 1995.

F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituta Italiano degli Attuari*, 4:221–424, 1933.

K. C. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J. J. Tyson. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Molecular Biology of the Cell*, 11(1):369–391, Jan. 2000.

K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak, and J. J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Molecular Biology of the Cell*, 15(8):3841–3862, Aug. 2004.

J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer, and D. Botstein. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73, May 1997.

S. Chowdhury, K. W. Smith, and M. C. Gustin. Osmotic stress and the yeast cytoskeleton: phenotype-specific suppression of an actin mutation. *The Journal of Cell Biology*, 118(3):561–71, Aug. 1992.

M. Cook and M. Tyers. Size control goes global. *Current Opinion in Biotechnology*, 18 (4):341–350, Aug. 2007.

N. A. Cookson, S. W. Cookson, L. S. Tsimring, and J. Hasty. Cell cycle-dependent variations in protein concentration. *Nucleic Acids Research*, 38(8):2676–2681, Dec. 2009.

S. Courbet, S. Gay, N. Arnoult, G. Wronka, M. Anglana, O. Brison, and M. Debatisse. Replication fork movement sets chromatin loop size and origin choice in mammalian cells. *Nature*, 455(7212):557–560, Sept. 2008.

F. Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, 1970.

D. M. Czajkowsky, J. Liu, J. L. Hamlin, and Z. Shao. DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI. *Journal of Molecular Biology*, 375(1):12–19, Jan. 2008.

C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Murray, John, London, 2 edition, 1860.

R. a. M. de Bruin, W. H. McDonald, T. I. Kalashnikova, J. Yates, and C. Wittenberg. Cln3 activates G1-specific transcription via phosphorylation of the SBF bound repressor Whi5. *Cell*, 117(7):887–98, June 2004.

H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, Jan. 2002.

A. De Moivre. *The Doctrine of Chances*, volume 2. Woodfall, 2 edition, 1738.

A. P. S. de Moura, R. Retkute, M. Hawkins, and C. a. Nieduszynski. Mathematical modelling of whole chromosome replication. *Nucleic Acids Research*, 6:1–11, May 2010.

A. Dershowitz and C. S. Newlon. The effect on chromosome stability of deleting replication origins. *Molecular and Cellular Biology*, 13(1):391–398, Jan. 1993.

A. Dershowitz, M. Snyder, M. Sbia, J. H. Skurnick, L. Y. Ong, and C. S. Newlon. Linear derivatives of Saccharomyces cerevisiae chromosome III can be maintained in the absence of autonomously replicating sequence elements. *Molecular and Cellular Biology*, 27(13):4652–4663, July 2007.

A. M. Deshpande and C. S. Newlon. DNA replication fork pause sites dependent on transcription. *Science*, 272(5264):1030–1033, May 1996.

P. Deuflhard and A. Hohmann. *Numerische Mathematik I*. Walter de Gruyter, 2 edition, 2002. ISBN 3110171821.

C. Dez and D. Tollervey. Ribosome synthesis meets the cell cycle. *Current Opinion in Microbiology*, 7(6):631–637, Dec. 2004.

S. Di Talia, J. M. Skotheim, J. M. Bean, E. D. Siggia, and F. R. Cross. The effects of molecular noise and size control on variability in the budding yeast cell cycle. *Nature*, 448(7156):947–951, Aug. 2007.

S. Di Talia, H. Wang, J. M. Skotheim, A. P. Rosebrock, B. Futcher, and F. R. Cross. Daughter-specific transcription factors regulate cell size control in budding yeast. *PLoS Biology*, 7(10):e1000221, Oct. 2009.

J. Diffley and K. Labib. The chromosome replication cycle. *Journal of Cell Science*, 115 (5):869, Mar. 2002.

M. Dobrzynski and F. J. Bruggeman. Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences USA*, 106(8):2583–2588, Feb. 2009.

A. Donaldson, M. Raghuraman, and K. Friedman. CLB5-dependent activation of late replication origins in S. cerevisiae. *Molecular Cell*, 2(2):173–182, Aug. 1998.

J. Donato, S. Chung, and B. Tye. Genome-wide hierarchy of replication origin usage in Saccharomyces cerevisiae. *PLoS Genetics*, 2(9):e141, 2006.

A. Dutta and S. P. Bell. Initiation of DNA replication in eukaryotic cells. *Annual Review of Cell Developmental Biology*, 13:293–332, 1997.

N. K. Egilmez, J. B. Chen, and S. M. Jazwinski. Preparation and partial characterization of old yeast cells. *Journal of Gerontology*, 45(1):B9–17, Jan. 1990.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, Aug. 2002.

J. M. Enserink and R. D. Kolodner. An overview of Cdk1-controlled targets and processes. *Cell Division*, 5:11, Jan. 2010.

C. B. Epstein and F. R. Cross. CLB5: a novel B cyclin from budding yeast with a role in S phase. *Genes & Development*, 6(9):1695–1706, Sept. 1992.

X. Escoté, M. Zapater, J. Clotet, and F. Posas. Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nature Cell Biology*, 6(10):997–1002, Oct. 2004.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan. 2007.

W. L. Fangman and B. J. Brewer. Activation of replication origins within yeast chromosomes. *Annual Review of Cell Biology*, 7:375–402, 1991.

W. Feng, D. Collingwood, M. E. Boeck, L. a. Fox, G. M. Alvino, W. L. Fangman, M. K. Raghuraman, and B. J. Brewer. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nature Cell Biology*, 8 (2):148–155, Feb. 2006.

B. M. Ferguson and W. L. Fangman. A position effect on the time of replication origin activation in yeast. *Cell*, 68(2):333–339, Jan. 1992.

P. Françon, D. Maiorano, and M. Méchali. Initiation of DNA replication in eukaryotes: questioning the origin. *FEBS Letters*, 452(1-2):87–91, June 1999.

L. I. Francis, J. C. W. Randell, T. J. Takara, L. Uchima, and S. P. Bell. Incorporation into the prereplicative complex activates the Mcm2-7 helicase for Cdc7-Dbf4 phosphorylation. *Genes & Development*, 23(5):643–654, Mar. 2009.

K. L. Friedman, J. D. Diller, B. M. Ferguson, S. V. Nyland, B. J. Brewer, and W. L. Fangman. Multiple determinants controlling activation of yeast replication origins late in S phase. *Genes & Development*, 10(13):1595–1607, July 1996.

C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, 4 edition, 2009. ISBN 9783540707127.

R. C. Gentleman, V. J. Carey, D. M. Bates, and Others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, Jan. 2007.

V. I. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituta Italiano degli Attuari*, 4:92–99, 1933.

A. Goelzer and V. Fromion. Bacterial growth rate reflects a bottleneck in resource allocation. *Biochimica et Biophysica Acta*, June 2011.

A. Goldar, H. Labit, K. Marheineke, and O. Hyrien. A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS One*, 3(8):e2919, Jan. 2008.

A. Goldbeter and D. E. Koshland. An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences USA*, 78(11):6840–6844, 1981.

A. I. Goranov, M. Cook, M. Ricicova, G. Ben-Ari, C. Gonzalez, C. Hansen, M. Tyers, and A. Amon. The rate of cell growth is governed by cell cycle stage. *Genes & Development*, 23(12):1408–1422, June 2009.

A. Goren and H. Cedar. Replicating by the clock. *Nature Reviews Molecular Cell Biology*, 4(1):25–32, Jan. 2003.

D. D. Hall, D. D. Markwardt, F. Parviz, and W. Heideman. Regulation of the Cln3-Cdc28 kinase by cAMP in Saccharomyces cerevisiae. *The EMBO journal*, 17(15): 4370–8, Aug. 1998.

J. Hamlin, L. Mesner, O. Lar, R. Torres, S. Chodaparambil, and L. Wang. A revisionist replicon model for higher eukaryotic genomes. *Journal of Cellular Biochemistry*, 105 (2):321–329, Oct. 2008.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan. 2000.

D. Hanahan and R. A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144 (5):646–674, Mar. 2011.

L. Hartwell and T. Weinert. Checkpoints: controls that ensure the order of cell cycle events. *Science*, 246(4930):629–634, 1989.

L. H. Hartwell and M. W. Unger. Unequal division in Saccharomyces cerevisiae and its implications for the control of cell division. *The Journal of Cell Biology*, 75(2 Pt 1): 422–35, Nov. 1977.

L. H. Hartwell, J. Culotti, J. R. Pringle, and B. J. Reid. Genetic control of the cell division cycle in yeast: a model to account fo the order of cell cycle events is deduced from the phenotypes of the yeast mutants. *Science*, 183:46–51, 1974.

M. A. Henson. Dynamic modeling of microbial cell populations. *Current Opinion in Biotechnology*, 14(5):460–467, Oct. 2003.

S.-I. Hiraga, A. Hagihara-Hayashi, T. Ohya, and A. Sugino. DNA polymerases alpha, delta, and epsilon localize and function together at replication forks in Saccharomyces cerevisiae. *Genes Cells*, 10(4):297–309, Apr. 2005.

M. A. Hjortso and J. E. Bailey. Transient responses of budding yeast populations. *Mathematical Biosciences*, 63(1):121–148, Feb. 1983.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

O. Hyrien and A. Goldar. Mathematical modelling of eukaryotic DNA replication. *Chromosome Research*, Nov. 2009.

O. Hyrien, K. Marheineke, and A. Goldar. Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *BioEssays*, 25(2):116–125, Feb. 2003.

T. Ideker and D. Lauffenburger. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology*, 21(6):255–262, June 2003.

B. P. Ingalls and H. M. Sauro. Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. *Journal of Theoretical Biology*, 222(1):23–36, May 2003.

L. P. Jackson, S. I. Reed, and S. B. Haase. Distinct mechanisms control the stability of the related S-phase cyclins Clb5 and Clb6. *Molecular and Cellular Biology*, 26(6): 2456–2466, Mar. 2006.

F. Jacob and S. Brenner. On the regulation of DNA synthesis in bacteria: the hypothesis of the replicon. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 256:298–300, Jan. 1963.

W. Jaeger. *Aristotle - Metaphysica*. Oxford University Press, 1957.

F. Jarre and J. Stoer. *Optimierung*. Springer, 2003. ISBN 3540435751.

N. Ji, R. C. Allshire, A. J. Klar, and S. I. Grewal. A role for DNA polymerase alpha in epigenetic control of transcriptional silencing in fission yeast. *EMBO Journal*, 20(11): 2857–2866, June 2001.

G. C. Johnston, J. R. Pringle, and L. H. Hartwell. Coordination of growth with cell division in the yeast Saccharomyces cerevisiae. *Experimental Cell Research*, 105(1): 79–98, Mar. 1977.

G. C. Johnston, C. W. Ehrhardt, A. Lorincz, and B. L. Carter. Regulation of cell size in the yeast Saccharomyces cerevisiae. *Journal of Bacteriology*, 137(1):1–5, Jan. 1979.

G. C. Johnston, R. a. Singer, S. O. Sharrow, and M. L. Slater. Cell Division in the Yeast Saccharomyces cerevisiae Growing at Different Rates. *Microbiology*, 118(2):479–484, June 1980.

P. Jorgensen, J. L. Nishikawa, B.-J. Breitkreutz, and M. Tyers. Systematic identification of pathways that couple cell growth and division in yeast. *Science*, 297(5580):395–400, July 2002.

P. Jorgensen, I. Rupes, J. R. Sharom, L. Schneper, J. R. Broach, and M. Tyers. A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes & Development*, 18(20):2491–505, Oct. 2004.

P. Jorgensen, N. P. Edgington, B. L. Schneider, I. Rupes, M. Tyers, and B. Futcher. The size of the nucleus increases as yeast cells grow. *Molecular Biology of the Cell*, 18(9): 3523–3532, Sept. 2007.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, Jan. 2000.

M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue):D354—-D357, Jan. 2006.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database issue):D480—-D484, Jan. 2008.

S. Kar, W. T. Baumann, M. R. Paul, and J. J. Tyson. Exploring the roles of noise in the eukaryotic cell cycle. *Proceedings of the National Academy of Sciences USA*, 106 (16):6471–6486, Apr. 2009.

K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5):491–6, Oct. 2003.

B. B. Kaufmann and A. van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetic Development*, 17(2):107–112, Apr. 2007.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, May 1983.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov. 2002.

E. Klipp, B. Nordlander, R. Krüger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nature Biotechnology*, 23(8):975–82, Aug. 2005.

E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. *Systems Biology: A Textbook*. Wiley-VCH, 1 edition, 2009. ISBN 978-3-527-31874-2.

H. Kohzaki, Y. Ito, and Y. Murakami. Context-dependent modulation of replication activity of Saccharomyces cerevisiae autonomously replicating sequences by transcription factors. *Molecular and Cellular Biology*, 19(11):7428–7435, Nov. 1999.

K. Koutroumpas and J. Lygeros. Modeling and analysis of DNA replication. *Automatica*, 47(6):1156–1164, June 2011.

C. Kühne and P. Linder. A new pair of B-type cyclins from Saccharomyces cerevisiae that function early in the cell cycle. *EMBO Journal*, 12(9):3437–3447, Sept. 1993.

T. Kunkel and P. Burgers. Dividing the workload at a eukaryotic replication fork. *Trends in Cell Biology*, 18(11):521–527, Nov. 2008.

J.-B. Lamarck. *Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux...* Dentu et L'Auteur, 2 edition, 1809.

S. Lanker, M. Valdivieso, and C. Wittenberg. Rapid degradation of the G1 cyclin Cln2 induced by CDK-dependent phosphorylation. *Science*, 271(5255):1597, 1996.

P.-S. Laplace. *Théorie analytique des probabilités*. Paris, Ve. Courcier, Paris, 1812.

N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, Aug. 2009.

M. Lee and P. Nurse. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc 2. *Nature*, 327(6117):31–35, 1987.

M. Lei, Y. Kawasaki, and B. K. Tye. Physical interactions among Mcm proteins and effects of Mcm dosage on DNA replication in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 16(9):5081–5090, Sept. 1996.

F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences USA*, 101(14):4781–4786, Apr. 2004.

R. Lucchini and J. M. Sogo. Chromatin structure and transcriptional activity around the replication forks arrested at the 3' end of the yeast rRNA genes. *Molecular and Cellular Biology*, 14(1):318–326, Jan. 1994.

J. Lygeros, K. Koutroumpas, S. Dimopoulos, I. Legouras, P. Kouretas, C. Heichinger, P. Nurse, and Z. Lygerou. Stochastic hybrid modeling of DNA replication across a complete genome. *Proceedings of the National Academy of Sciences USA*, 105(34): 12295–300, Aug. 2008.

D. M. MacAlpine, H. K. Rodríguez, and S. P. Bell. Coordination of replication and transcription along a Drosophila chromosome. *Genes & Development*, 18(24):3094–3105, Dec. 2004.

Y. Marahrens and B. Stillman. A yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science*, 255(5046):817–823, Feb. 1992.

H. J. McCune, L. S. Danielson, G. M. Alvino, D. Collingwood, J. J. Delrow, W. L. Fangman, B. J. Brewer, and M. K. Raghuraman. The temporal program of chromosome replication: genomewide replication in clb5{Delta} Saccharomyces cerevisiae. *Genetics*, 180(4):1833–1847, Dec. 2008.

M. Mechali. DNA replication origins: from sequence specificity to epigenetics. *Nature Reviews Genetics*, 2(8):640–645, Aug. 2001.

L. Michaelis and M. Menten. Die Kinetik der Invertinwirkung. *Biochemische Zeitung*, 49:333–369, 1913.

J. Mitchison. The growth of single cells II. Saccharomyces cerevisiae. *Experimental Cell Research*, 15(1):214–221, Aug. 1958.

J. M. Mitchison. *The biology of the cell cycle.* Cambridge University Press, Cambridge, 1971.

J. M. Mitchison. Growth during the cell cycle. *International Review of Cytology*, 226: 165–258, Jan. 2003.

D. Molenaar, R. van Berlo, D. de Ridder, and B. Teusink. Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular Systems Biology*, 5:323, 2009.

C. Moles, P. Mendes, and J. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13(11):2467–2474, 2003.

D. O. Morgan. Principles of CDK regulation. *Nature*, 374(6518):131–4, Mar. 1995.

S. Mori and K. Shirahige. Perturbation of the activity of replication origin by meiosis-specific transcription. *Journal of Biological Chemistry*, 282(7):4447–4452, Feb. 2007.

I. Mura and A. Csikász-Nagy. Stochastic Petri Net extension of a yeast cell cycle model. *Journal of Theoretical Biology*, 254(4):850–860, Oct. 2008.

D. B. Murray, M. Beckmann, and H. Kitano. Regulation of yeast oscillatory dynamics. *Proceedings of the National Academy of Sciences USA*, 104(7):2241–6, Feb. 2007.

C. S. Newlon and J. F. Theis. The structure and function of yeast ARS elements. *Current Opinion in Genetic Development*, 3(5):752–758, Oct. 1993.

C. S. Newlon, L. R. Lipchitz, I. Collins, A. Deshpande, R. J. Devenish, R. P. Green, H. L. Klein, T. G. Palzkill, R. B. Ren, and S. Synn. Analysis of a circular derivative of Saccharomyces cerevisiae chromosome III: a physical map and identification and location of ARS elements. *Genetics*, 129(2):343–357, Oct. 1991.

V. Q. Nguyen, C. Co, K. Irie, and J. J. Li. Clb/Cdc28 kinases promote nuclear export of the replication initiator proteins Mcm2-7. *Current Biology*, 10(4):195–205, Feb. 2000.

S. A. Nick-McElhinny, D. A. Gordenin, C. M. Stith, P. M. J. Burgers, and T. A. Kunkel. Division of labor at the eukaryotic replication fork. *Molecular Cell*, 30(2):137–144, Apr. 2008.

C. A. Nieduszynski, J. J. Blow, and A. D. Donaldson. The requirement of yeast replication origins for pre-replication complex proteins is modulated by transcription. *Nucleic Acids Research*, 33(8):2410–2420, 2005.

C. A. Nieduszynski, Y. Knox, and A. D. Donaldson. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes & Development*, 20(14): 1874–1879, July 2006.

C. A. Nieduszynski, S. Hiraga, P. Ak, C. J. Benham, and A. D. Donaldson. OriDB: a DNA replication origin database. *Nucleic Acids Research*, 35(Database issue):D40—-D46, Jan. 2007.

D. Noble. *The Music of Life: Biology beyond genes*. Oxford University Press, 2008. ISBN 0199228361.

P. Nurse. A long twentieth century of the cell cycle and beyond. *Cell*, 100(1):71–78, Jan. 2000.

T. Ogawa and T. Okazaki. Discontinuous DNA replication. *Annual Review of Biochemistry*, 49:421–457, 1980.

R. Okazaki, T. Okazaki, and K. Sakabe. Mechanism of DNA replication possible discontinuity of DNA chain growth. *Japanese Journal of Medical Science and Biology*, 20(3):255–260, June 1967.

P. Pasero, A. Bensimon, and E. Schwob. Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. *Genes & Development*, 16(19):2479–2484, Oct. 2002.

P. K. Patel, B. Arcangioli, S. P. Baker, A. Bensimon, and N. Rhind. DNA replication origins fire stochastically in fission yeast. *Molecular Biology of the Cell*, 17(1):308–316, Jan. 2006.

M. Peter and I. Herskowitz. Direct inhibition of the yeast cyclin-dependent kinase Cdc28-Cln by Far1. *Science*, 265(5176):1228–31, Aug. 1994.

L. Petzold. Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations. *SIAM Journal on Scientific and Statistical Computing*, 4(1):136–148, 1983.

T. Pfeiffer, S. Schuster, and S. Bonhoeffer. Cooperation and competition in the evolution of ATP-producing pathways. *Science*, 292(5516):504–7, Apr. 2001.

J. Pines. Cyclins and cyclin-dependent kinases: a biochemical view. *The Biochemical Journal*, 308(3):697–711, June 1995.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database issue):D61—-D65, Jan. 2007.

R Development Core Team. R: A Language and Environment for Statistical Computing. 2007.

M. Raghuraman and B. Brewer. Molecular analysis of the replication program in unicellular model organisms. *Chromosome Research*, 18(1):19–34, Jan. 2010.

M. K. Raghuraman, E. a. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman. Replication dynamics of the yeast genome. *Science*, 294(5540):115–121, Oct. 2001.

A. Ramanathan and S. Schreiber. Multilevel regulation of growth rate in yeast revealed using systems biology. *Journal of Biology*, 6(2):3, 2007.

N. Rhind. DNA replication timing: random thoughts about origin firing. *Nature Cell Biology*, 8(12):1313–1316, Dec. 2006.

N. Rhind, S. Yang, and J. Bechhoefer. Reconciling stochastic origin firing with defined replication timing. *Chromosome Research*, 18(1):35–43, Nov. 2010.

C. Rivin and W. Fangman. Replication fork rate and origin activation during the S phase of Saccharomyces cerevisiae. *The Journal of Cell Biology*, 85(1):108–115, 1980.

O. Rokhlenko, T. Shlomi, R. Sharan, E. Ruppin, and R. Y. Pinter. Constraint-based functional similarity of metabolic genes: going beyond network topology. *Bioinformatics*, 23(16):2139–2146, Aug. 2007.

D. Rudra and J. R. Warner. What better measure than ribosome synthesis? *Genes & Development*, 18(20):2431–6, Oct. 2004.

I. Rupes. Checking cell size in yeast. *Trends in Genetics*, 18(9):479–85, Sept. 2002.

M. Sabouri-Ghomi, A. Ciliberto, S. Kar, B. Novak, and J. J. Tyson. Antagonism and bistability in protein interaction networks. *Journal of Theoretical Biology*, 250(1): 209–218, Jan. 2008.

A. Sackmann, M. Heiner, and I. Koch. Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7:482, Jan. 2006.

B. Schneider, J. Zhang, J. Markwardt, G. Tokiwa, T. Volpe, S. Honey, and B. Futcher. Growth rate and cell size modulate the synthesis of, and requirement for, G1-phase cyclins at start. *Molecular and Cellular Biology*, 24(24):10802, 2004.

E. Schwob and K. Nasmyth. CLB5 and CLB6, a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae. *Genes & Development*, 7(7A):1160–1175, July 1993.

M. D. Sekedat, D. Fenyö, R. S. Rogers, A. J. Tackett, J. D. Aitchison, and B. T. Chait. GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. *Molecular Systems Biology*, 6(353):353, Jan. 2010.

SGD Project. Saccharomyces Genome Database. URL `http://downloads.yeastgenome.org/chromosomal_feature/03.03.2010`.

K. Shirahige, T. Iwasaki, M. B. Rashid, N. Ogasawara, and H. Yoshikawa. Location and characterization of autonomously replicating sequences from chromosome VI of Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 13(8):5043–56, Aug. 1993.

J. M. Skotheim, S. Di Talia, E. D. Siggia, and F. R. Cross. Positive feedback of G1 cyclins ensures coherent cell cycle entry. *Nature*, 454(7202):291–296, July 2008.

H. Soueidan, D. Sherman, and M. Nikolski. BioRica: A multi model description and simulation system. *F0SBE*, pages 279–287, 2007.

C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–97, Dec. 1998.

T. W. Spiesser and E. Klipp. Different Groups of Metabolic Genes Cluster Around Early and Late Firing Origins of Replication in Budding Yeast. *Genome Informatics*, 24: 179–192, 2010.

T. W. Spiesser, E. Klipp, and M. Barberis. A model for the spatiotemporal organization of DNA replication in Saccharomyces cerevisiae. *Molecular Genetics and Genomics*, 282(1):25–35, July 2009.

T. W. Spiesser, C. Diener, M. Barberis, and E. Klipp. What Influences DNA Replication Rate in Budding Yeast? *PLoS One*, 5(4):e10203, 2010.

B. Stillman. Origin recognition and the chromosome cycle. *FEBS Letters*, 579(4):877–884, Feb. 2005.

D. T. Stinchcomb, K. Struhl, and R. W. Davis. Isolation and characterisation of a yeast chromosomal replicator. *Nature*, 282(5734):39–43, Nov. 1979.

D. Stirzaker. *Stochastic Processes and Models*. Oxford University Press, 2005. ISBN 978-0-19-856813-1.

Z. Szallasi, J. Stelling, and V. Periwal, editors. *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*. The MIT Press, 2006. ISBN 0262195488.

A. Tabancay and Others. Eukaryotic DNA replication in a chromatin context. *Current Topics in Developmental Biology*, 76:129–184, 2006.

M. Takahashi. A model for the spatio-temporal organization of DNA replication in mammalian cells. *Journal of Theoretical Biology*, 129(1):91–115, Nov. 1987.

D. Y. Takeda and A. Dutta. DNA replication and progression through S phase. *Oncogene*, 24(17):2827–2843, Apr. 2005.

J. F. Theis and C. S. Newlon. The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. *Proceedings of the National Academy of Sciences USA*, 94(20):10786–10791, Sept. 1997.

C. B. Tyson, P. G. Lord, and A. E. Wheals. Dependency of size of Saccharomyces cerevisiae cells on growth rate . *Journal of Bacteriology*, 138(1):92–98, 1979.

J. J. Tyson and B. Novak. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *Journal of Theoretical Biology*, 210(2):249–63, May 2001.

A. J. van Brabant, C. D. Buchanan, E. Charboneau, W. L. Fangman, and B. J. Brewer. An origin-deficient yeast artificial chromosome triggers a cell cycle checkpoint. *Molecular Cell*, 7(4):705–713, Apr. 2001.

G. van Rossum. *Python reference manual*. CWI (Centre for Mathematics and Computer Science), Amsterdam, The Netherlands, The Netherlands, 1995.

C. M. Waage P.; Gulberg. Studies concerning affinity. *Journal of Chemical Education*, 63:1044, 1986.

J. R. Warner. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Science*, 24(11):437–440, Nov. 1999.

M. Weinreich, M. a. Palacios DeBeer, and C. a. Fox. The activities of eukaryotic replication origins in chromatin. *Biochimica et Biophysica Acta*, 1677(1-3):142–157, Mar. 2004.

R. E. Wellinger, F. Prado, and A. Aguilera. Replication fork progression is impaired by transcription in hyperrecombinant yeast cells lacking a functional THO complex. *Molecular and Cellular Biology*, 26(8):3327–3334, Apr. 2006.

H. V. Westerhoff, C. Winder, H. Messiha, E. Simeonidis, M. Adamczyk, M. Verma, F. J. Bruggeman, and W. Dunn. Systems biology: the elements and principles of life. *FEBS Letters*, 583(24):3882–3890, Dec. 2009.

E. Wintersberger. Why is there late replication? *Chromosoma*, 109(5):300–307, 2000.

J. J. Wyrick, J. G. Aparicio, T. Chen, J. D. Barnett, E. G. Jennings, R. a. Young, S. P. Bell, and O. M. Aparicio. Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science*, 294(5550): 2357–2360, Dec. 2001.

L. Xiao and A. Grove. Coordination of Ribosomal Protein and Ribosomal RNA Gene Expression in Response to TOR Signaling. *Current Genomics*, 10(3):198–205, May 2009.

W. Xu, J. G. Aparicio, O. M. Aparicio, and S. Tavaré. Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in S. cerevisiae. *BMC Genomics*, 7:276, Jan. 2006.

N. Yabuki and H. Terashima. Mapping of early firing origins on a replication profile of budding yeast. *Genes to Cells*, 7(8):781–789, Aug. 2002.

M. Yamashita, Y. Hori, T. Shinomiya, C. Obuse, T. Tsurimoto, H. Yoshikawa, and K. Shirahige. The efficiency and timing of initiation of replication of multiple replicons of Saccharomyces cerevisiae chromosome VI. *Genes to Cells*, 2(11):655–665, Nov. 1997.

S. Yang and J. Bechhoefer. How Xenopus laevis embryos replicate reliably: Investigating the random-completion problem. *Physical Review E*, 78(4):1–15, Oct. 2008.

S. C.-H. Yang, N. Rhind, and J. Bechhoefer. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Molecular Systems Biology*, 6(404):1–13, Aug. 2010.

M. Zannis-Hadjopoulos and G. Price. Regulatory parameters of DNA replication. *Critical Reviews in Eukaryotic Gene Expression*, 8(1):81, 1998.

Y. Zhang, M. Qian, Q. Ouyang, M. Deng, F. Li, and C. Tang. Stochastic model of yeast cell-cycle network. *Physica D: Nonlinear Phenomena*, 219(1):35–39, July 2006.

Z. Zhang, K. Shibahara, and B. Stillman. PCNA connects DNA replication to epigenetic inheritance in yeast. *Nature*, 408(6809):221–225, Nov. 2000.

L. P. Zhao, R. Prentice, and L. Breeden. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proceedings of the National Academy of Sciences USA*, 98(10):5631–5636, May 2001.

J. Zhou, C. Chau, Z. Deng, W. Stedman, and P. M. Lieberman. Epigenetic control of replication origins. *Cell Cycle*, 4(7):889–892, July 2005.

C. Zhu, R. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997.

L. Zou and B. Stillman. Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at replication origins controlled by S-phase cyclin-dependent kinases and Cdc7p-Dbf4p kinase. *Molecular and Cellular Biology*, 20(9):3086–3096, May 2000.

List of Figures

List of Tables

Acknowledgments

Here, I would like to thank everybody who has contributed to either my thesis or my well-being in the last three years.

First of all, I am deeply indebted to my supervisor **Edda Klipp**. She has guided my scientific education for the past years and gave me the opportunity to meet many people from the scientific community from all over the world.

Second, I thank **Matteo Barberis** for many years of support and fruitful teamwork on the exciting DNA replication project.

I am also grateful to **Marcus Krantz**, who offered me guidance and vital comments within the size control project. He has engaged me in many challenging discussion about the nature of cellular biology and how best to display it.

I also want to express my gratitude to **Christian Diener** and **Max Flöttman** for inspiration, collaboration, technical support, uncountable discussions, constant encouragement and finally, at times, distraction.

Many provided valuable last minute suggestions and helped proofreading, which undoubtedly improved the quality of my thesis. These were, unless already mentioned, my friends **Max Wend** and **Julian Henneberg** and my colleagues **Judith Wodke** and **Clemens Kühn**.

Furthermore, I would like to thank the entire **Group of Theoretical Biophysics** at the Humboldt-Universität zu Berlin for kind assistance throughout the formulation of this thesis and also the International Research Training Group (**IRTG**) for exciting retreats, journeys and for funding.

I especially would like to mention **Janek Hermann-Friede**, **Felipe Ceballos**, **Jörn Dietrich**, **Phillip Schmeling**, **Serkan Altuglu**, **Naomi Ryland** and **Andreas Werner**, because their friendship and support was an immeasurable factor in this intense time.

All my life, I could rely on my family **Marion Spiesser-Grunow**, **Wolm Spiesser**, **Natalie Spiesser**, **Ronja Spiesser** and lately also my little baby-nephew **Jacob**. Their love, kindness and encouragements are the foundation of my well-being and success.

Last but not least, I would like to thank **Julika Schmitz** for her love, company and care. She has accompanied me in every step of my work in these last months and gave me feedback and valuable comments in countless discussions.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 02.08.2011 Thomas W. Spießer

Publications

Thomas W. Spießer

T. W. Spiesser and E. Klipp.
Different groups of metabolic genes cluster around early and late firing origins of replication.
Genome Informatics 24:179-192, 2010.

M. Barberis, T. W. Spiesser and E. Klipp.
Kinetic modelling of DNA replication initiation in budding yeast.
Genome Informatics 24:1-20, 2010.

T. W. Spiesser, C. Diener, M. Barberis and E. Klipp.
What influences DNA replication rate in budding yeast?
PLoS One 5:e10203, 2010.

M. Barberis, T. W. Spiesser and E. Klipp.
Replication origins and timing of temporal replication in budding yeast: how to solve the conundrum.
Current Genomics 11:199-211, 2010.

T. W. Spiesser, E. Klipp and M. Barberis.
A model for the spatiotemporal organization of DNA replication in *Saccharomyces cerevisiae*.
Molecular Genetics and Genomics 282:25-35, 2009.

Berlin, 02.08.2011